Development of clustering algorithms for categorical data and applications in Health

Abdoul Jalil Djiberou Mahamadou

UNIVERSITY CLERMONT AUVERGNE

DOCTORAL SCHOOL
SCIENCES FOR ENGINEER OF CLERMONT-FERRAND

THESIS

Defended by

DJIBEROU MAHAMADOU Abdoul Jalil

In order to become

DOCTOR OF UNIVERSITY

SPECIALITY : COMPUTER SCIENCE

Development of clustering algorithms for categorical data and applications in Health

Defended on September 17th 2021

Jury :

Pr.	Vincent Barra	Université Clermont Auvergne	France	President
Pr.	Ana Belén Ramos-Guajardo	Universidad de Oviedo	Spain	Examiner
Pr.	Manuel Ojeda Aciego	Universidad de Málaga	Spain	Examiner
HDR Dr	Marie-Jeanne Lesot	Sorbonne Université	France	Reviewer
HDR Dr	Benjamin Quost	Université Technologique de Compiène	France	Reviewer
Pr.	Mephu Nguifo Engelbert	Université Clermont Auvergne	France	Director of thesis
Dr	Antoine Violaine	Université Clermont Auvergne	France	Supervisor
Pr.	Moreno Sylvain	Simon Fraser University	Canada	Supervisor

Acknowledgements

During my three years of the thesis, I have been working with awesome people that helped and supported me to make this adventure a success. I would like to dedicate the following words to these people.

My acknowledgements go first to my thesis supervisors Dr Antoine Violaine, Pr. Mephu Nguifo Engelbert, and Pr. Moreno Sylvain. I have a great chance to be one of your PhD students. Since the starting of this thesis, you have been providing me with all the resources I needed. Your advice guided me in each of the steps of my thesis. Thank you for your trust, your thoroughness, and your simplicity which helped me to be the best of myself.

I would like to thank all the members of my committee thesis and the reviewers of my published papers for their feedback that helped me to improve my work.

My special thanks go to HDR Dr Benjamin Quost and HDR Dr Marie-Jeanne Lesot who accepted to review this report, to the examiners Pr. Manuel Ojeda Aciego and Pr. Ana Belén Ramos-Guajardo, and finally to the president of the jury Pr. Vincent Barra for the evaluation of my work.

My acknowledgements also go to all my colleagues at LIMOS and The Digital Health Research Laboratory in Canada particularly, to Emma Rodrigues, Vasily Vakorin, Greg Christie, and all the MINERS workgroup members of LIMOS for the fruitful discussions and their continuous support during these three years. My sincere thanks go to Beatrice Bourdieu and Allegranzini Damien for their administrative supports.

Finally, I would like to thank my family who trusted in me and never failed to support me.

Abstract

Clustering is a popular unsupervised machine learning method that consists of grouping similar data objects in the same group and dissimilar objects in different groups. Among the clustering family methods, we can distinguish the partition-based methods which produce partitions of data objects. The type of the obtained partitions depends on the theory used. Hard partitions can be obtained with the hard sets theory whereas imprecision and uncertainty theories such as the fuzzy sets theory and the Dempster-Shafer theory of evidence can be used to obtain fuzzy partitions.

In this thesis, an extension of the fuzzy k-modes clustering method referred to as *categorical fuzzy entropy c-means* is firstly proposed. The new method uses the fuzzy sets theory to model the imprecision of object assignments to clusters and the representation of the centers of the clusters by associating weights to each attribute category which indicates their importance. A second new method referred to as *categorical evidential c-means* is proposed as a categorical version the *evidential c-means*. The latter method uses the Dempster-Shafer theory of evidence to capture the uncertainty of object labeling.

Several experiments on different datasets were conducted to illustrate the strengths of the new methods and to compare them with existing numerical and categorical clustering methods. In addition, the two methods were used to investigate the replication of new findings in developmental sciences on the influence of lifestyle factors on the cognitive health. The results of these experiments showed that the proposed methods have good performances and can handle imperfect data. Finally, research directions are given to extend the two methods to capture non-linear relationships among the variables of the input data and to suit time series data.

Key words: categorical data, fuzzy clustering, fuzzy centers, belief functions, imprecision, uncertainty, health.

Résumé

La classification non supervisée est une méthode d'apprentissage automatique populaire qui consiste à regrouper des objets de données similaires dans le même groupe et des objets dissemblables dans différents groupes. Parmi les méthodes de classification, on peut distinguer les méthodes basées sur des partitions qui produisent des partitions d'objets de données. Selon la théorie utilisée, les partitions obtenues peuvent être de différents types. En utilisant la théorie des ensembles (durs), les partitions produite sont dites dures. Les théories d'imprécision et d'incertitude telles que la théorie des ensembles flous et la théorie des fonctions de croyances de Dempster-Shafer peuvent être utilisées pour obtenir des partitions floues.

Dans cette thèse, une extension de la méthode de classification des k-modes flous appelée *c-moyennes floues catégorielles avec entropie* est proposée dans un premier temps. La nouvelle méthode utilise la théorie des ensembles flous pour modéliser l'imprécision des affectations d'objets aux classes et la représentations des centres des classes en associant des poids à chaque catégorie d'attributs qui indiquent leur importance. Par la suite, une deuxième nouvelle méthode appelée *c-moyennes évidentielles catégorielles* est proposée comme une version catégorielle de l'algorithme des *c-moyennes évidentielles*. Cette dernière méthode utilise la théorie des fonctions de croyance de Dempster-Shafer afin de modéliser l'incertitude de la classification des objets.

Plusieurs expériences sur différentes données ont été menées pour illustrer les points forts des nouvelles méthodes et pour comparer ces dernières avec des méthodes de classification numériques et catégorielles existantes. En outre, les deux méthodes ont été utilisées pour étudier la réplication de nouvelles découvertes en sciences du développement sur l'influence des facteurs liés au mode de vie sur la santé cognitive. Les résultats de ces expériences ont montré que les méthodes proposées ont de bonnes performances et peuvent gérer des données imparfaites. Enfin, des orientations de recherche sont données pour étendre les deux méthodes afin de capturer les relations non linéaires entre les variables des données d'entrée et pour des données de temporelles.

Mots clés: données catégorielles, classification floue, centres flous, fonctions de croyance, imprécision, incertitude, santé.

Publications

Thesis publications

International conferences

1. A. J. Djiberou Mahamadou, V. Antoine, E. M. Nguifo and S. Moreno, "Categorical fuzzy entropy c-means" 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, UK.

2. A. J. Djiberou Mahamadou, V. Antoine, G. J. Christie, and S. Moreno, "Evidential clustering for categorical data," 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, LA, USA, 2019.

National conferences

1. A. J. Djiberou Mahamadou, V. Antoine, E. M. Nguifo and S. Moreno, "Apport de l'entropie pour les c-moyennes floues sur des données catégorielles", EGC 2021.

Other publications

International journals (Submitted paper)

1. "Pilot study of eDOL, a new mHealth application and web platform for medical follow-up of chronic pain patients", KERCKHOVE, Delage, Cambier, Cantagrel, Serra, Marcaillou, Maindet-Dominici, PICARD, Martiné, Deleens, TROUVIN, Fourel, Espagne-Dubreuilh, Douay, FOULON, DUFRAISSE, GOV, Viel, JEDRYKA, Pouplin, Lestrade, Combe, Perrot, Perocheau, De Brisson, Vergne-Salle, Mertens, Pereira, **Djiberou Mahamadou**, Antoine, Corteval, Eschalier, Dualé, Attal, Authier; Journal of Medical Internet Research.

Contents

I Background **20**

1 Imprecision and uncertainty theories **21**
1.1 Hard sets theory as the foundation of imprecision and uncertainty theories . . . 22
1.2 Probability theory . . . 25
1.3 Fuzzy sets theory . . . 29
1.4 Theory of evidence . . . 32

2 Clustering algorithms **44**
2.1 Types of clustering partitions . . . 46
2.2 Partitioning algorithms . . . 52

II Contributions **70**

3 Categorical fuzzy entropy c-means **71**
3.1 Issues in *FC* . . . 72
3.2 New updates of weights in *FC* algorithm . . . 72
3.3 *CFE*: fuzzy entropy c-means . . . 76
3.4 Experiments . . . 79
3.5 Strengths of *CFE* . . . 92
3.6 Limitations of *CFE* . . . 93

4 Categorical evidential c-means **99**
4.1 The need for subsets of clusters . . . 100
4.2 Fuzzy centers and distance . . . 101
4.3 cat-ECM: categorical evidential c-means . . . 102
4.4 Experiments . . . 105
4.5 Strengths of cat-ECM . . . 111
4.6 Limitations of cat-ECM . . . 117

III Applications **120**

5 Differential susceptibility in older adults: cluster analysis perspectives **121**
5.1 Background . . . 122
5.2 Experiments . . . 127
5.3 Discussions . . . 142
5.4 Conclusion . . . 142

A Complementary experiments results on CFE **148**
A.1 CFE scores . . . 148
A.2 Critical difference diagrams . . . 150

B Complementary experiments results on cat-ECM **154**
B.1 cat-ECM scores . . . 154
B.2 Critical difference diagrams . . . 157
B.3 Partitions comparisons . . . 158

C Complementary real-world applications results **161**
C.1 Descriptive statistics on the HRS dataset . . . 161
C.2 Frequencies of cognitive categories in clusters . . . 162
C.3 Test values in each cognitive categories . . . 163

Nomenclature

β	Fuzziness coefficient in clustering algorithms.
$\sqsupset$	Triangular norm.
$\cap$	Intersection operator.
$\cup$	Union operator.
$\urcorner$	Triangular co-norm.
δ	Relative distance
ϵ	Input threshold used in the convergence of clustering algorithms.
γ	Statistical test level.
$\mathcal{A}$	Fuzzy set.
$\mathcal{J}$	Overall number of categories in $\mathbf{X}$, i.e., $\mathcal{J} = \sum_{l=1}^{p} n_l$.
$\mathcal{L}$	Lagrangian.
$\mathrm{A_l}$	l_{th} attribute of $\mathbf{X}$.
$\mathbf{A}$	Hard set.
$\mathbf{X}$	Space of objects, if finite, denoted by $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_i}, \ldots, \mathbf{x_n}\}$.
$\mu_{\mathcal{A}}$	Membership degree function of $\mathcal{A}$.
μ_{ik}	Fuzzy membership degree of object i in the k_{th} cluster.
$\neg$	Complement of a set.
$\mathbf{\Omega}$	Space of finite hard sets. In evidence theory, $\mathbf{\Omega}$ is called frame of discernment.
ω	Subset of $\mathbf{\Omega}$.
$\oplus$	Combination operator.

Π	Subset of *cat-ECM* clusters containing outliers.
Ψ	Fuzzy entropy coefficient of *CFE*.
ρ	Distance to the outliers cluster in cat-ECM.
τ	Algorithms iterations number.
ϱ	Number of desired subsets of clusters in *ECM* and *cat-ECM* algorithm. For instance, $\varrho = c + 2$ or $\varrho = 2^c$.
$a_l^{(t)}$	t_{th} category of attribute $\mathrm{A_l}$.
bel	Belief function.
c	Number of clusters.
H	Shannon's entropy.
$m(*)$	Mass function associated to $\mathbf{A}$.
N	Nonspecificity.
n	Number of objects in $\mathbf{X}$.
n_l	Number of categories in A_l.
p	Number of attributes.
pl	Plausibility function.
Q	Time complexity to solve the linear system in equation (2.20).
T	Number of iterations of an algorithm until convergence.
$w_{kl}^{(t)}$	Weight associated to the t_{th} category of the l_{th} attribute in the k_{th} cluster.
M	Evidential partition of X.
U	Fuzzy partition of X.
V	Fuzzy centers.
BetP	Belief to probability transformation function.
cat-ECM	Categorical evidential c-means.
CFE	Categorical fuzzy entropy c-means.
CRI	Credal Rand Index.
ECM	Evidential c-means.

FC Fuzzy centers clustering.

FC* Fuzzy centers clustering with hard centers.

FCM Fuzzy c-means.

FKM Fuzzy k-modes.

FRI Fuzzy Rand Index.

FS Fuzzy Silhouette index.

KM K-means.

N Nonspecificity.

PC Partition Coefficient.

PE Partition Entropy.

RI Rand Index.

List of Figures

3.1 PCA results on datasets. The inertia on the axes indicates the amount of explained variance. 83
3.2 Experimental protocol where D_h are datasets with $h = 1, \ldots, 9$. 87
3.3 Critical difference diagram obtained for PC scores with $\gamma = 0.1$ and $\beta = 1.5$. 90
3.4 Critical difference diagram obtained for PE scores with $\gamma = 0.1$ and $\beta = 1.5$. 90
3.5 Critical difference diagram obtained for FRI scores with $\gamma = 0.2$ and $\beta = 1.9$. 91
3.6 Critical difference diagram obtained for RI scores with $\gamma = 0.25$ and $\beta = 2$. 91
3.7 Critical difference diagram obtained for FS scores with $\gamma = 0.7$ and $\beta = 1.2$. 92
3.8 Simulated categorical data with $p = 0.8$. 93

4.1 Ordinal categorical version of the Butterfly dataset. 101
4.2 Critical difference diagram obtained for PC and PE scores with $\gamma = 0.1$ and $\beta = 1.1$. 109
4.3 Critical difference diagram obtained PC scores with $\gamma = 0.1$ and $\beta = 1.2$. 109
4.4 Critical difference diagram obtained PE scores with $\gamma = 0.1$ and $\beta = 1.2$. 110
4.5 Critical difference diagram obtained FRI scores with $\gamma = 0.5$ and $\beta = 1.1$. 110
4.6 Nonspecificity against Consistency obtained on the Zoo dataset with $\beta = 1.1$. 112
4.7 Nonspecificity against Consistency obtained on the Votes dataset with $\beta = 1.1$. 113
4.8 Nonspecificity against Consistency obtained on the Credits dataset with $\beta = 1.1$. 114
4.9 Mass obtained with *cat-ECM* on the ordinal categorical Butterfly dataset. 115
4.10 Subsets of clusters obtained with *cat-ECM* on the Zoo dataset. 116

5.1 Marginal effects analysis from [1] on the most significant lifestyle factors from the OLR model (see subsection 5.2.1 for a description of each factor). The x-axis corresponds to the 5 cognitive categories and the y-axis to the marginal scores. . 125
5.2 Projection of the HRS data with principal component analysis with different cognitive categories configurations. The x and y axes respectively correspond to the first and second components which explain 19% and 12% of the total inertia (variance). 131
5.3 Relative and global percentage of cognitive categories in merged clusters from *CFE*. 134
5.4 Relative and global percentage of cognitive categories in merged clusters from *cat-ECM*. 134
5.5 Test values of lifestyle factors categories in each cognitive category where the classes {12} and {45} correspond respectively to the merge of cognitive categories 1 and 2 and 4 and 5. The class {3} corresponds to cognitive category 3. 136
5.6 test values of lifestyle factors categories obtained from the hard partition of *CFE*. The classes {14} and {235} correspond respectively to the merging of clusters 1, 4 and 2, 3, 5 and correspond to older adults respectively with low and high cognitive levels. 137
5.7 test values of lifestyle factors categories obtained from the hard partition of *cat-ECM*. The classes {12} and {345} correspond respectively to the merging of clusters 1, 2 and 3, 4, 5 and correspond to older adults respectively with high and low cognitive levels. 138
5.8 Distribution of cognitive categories in focal sets obtained from *cat-ECM*. 139
5.9 Venn diagram from the focal sets obtained from *cat-ECM*. The numbers correspond to the number of objects in each subset. 140
5.10 Test values of focal sets obtained from *cat-ECM*. 141

A.1 Critical difference diagram obtained for PC scores with $\gamma = 0.1$ and $\beta = 1.6$. 152
A.2 Critical difference diagram obtained for PC scores with $\gamma = 0.1$ and $\beta = 1.7$. 152
A.3 Critical difference diagram obtained for PC scores with $\gamma = 0.1$ and $\beta = 1.8$. 152
A.4 Critical difference diagram obtained for PC scores with $\gamma = 0.1$ and $\beta = 1.9$. 152
A.5 Critical difference diagram obtained for PE scores with $\gamma = 0.1$ and $\beta = 1.6$. 153

A.6 Critical difference diagram obtained for PE scores with $\gamma = 0.1$ and $\beta = 1.7$. 153
A.7 Critical difference diagram obtained for FS scores with $\gamma = 0.7$ and $\beta = 1.8$. 153

B.1 Critical difference diagram obtained for RI scores with $\gamma = 0.1$ and $\beta = 2$. 157
B.2 Critical difference diagram obtained for RI scores with $\gamma = 0.1$ and $\beta = 1.1$. 157
B.3 Critical difference diagram obtained for FS scores with $\gamma = 0.1$ and $\beta = 1.1$. 157
B.4 Critical difference diagram obtained for FS scores with $\gamma = 0.1$ and $\beta = 2$. 158
B.5 Nonspecificity against Consistency obtained on the Zoo dataset with $\beta = 2$. 158
B.6 Nonspecificity against consistency obtained on the Votes dataset with $\beta = 2$. 159
B.7 Nonspecificity against consistency obtained on the Credits dataset with $\beta = 2$. 160

C.1 Relative and absolute cognitive categories frequencies in clusters from *CFE*. 162
C.2 Relative and absolute cognitive categories frequencies in clusters from *cat-ECM*. 163
C.3 vTests of lifestyle factors categories in each cognitive categories where the classes {12} and {45} correspond respectively the merged of cognitive categories 1 and 2 and 4 and 5. The class {3} corresponds to the cognitive category 3. . . . 164

List of Tables

1.1 Probability of failure of the vaccines on all the studied virus of COVID. . . . 36
1.2 Belief of failure of all vaccines on the variants. . . . 37
1.3 Plausibility of failure of all vaccines on the variants. . . . 37
1.4 Summary of fuzzy sets, and belief functions advantages, limitations, and properties. . . . 42

2.1 Matrix representation of the hard 2-partitions in Example 2.1.1 47
2.2 Matrix representation of the fuzzy 2-partition. . . . 49
2.3 Hard 2-partition generated from the fuzzy partition Table 2.2. 49
2.4 Example of an evidential partition. . . . 50
2.5 Categorical data to *one-hot encode*. . . . 60
2.6 *One-hot encoding* of Table 2.5. . . . 61
2.7 Categorical data set to illustrate fuzzy centers. . . . 65
2.8 Time and memory complexity of *KM*, *FCM*, *ECM*, *k-modes*, *FKM* and *FC* algorithms. . . . 68

3.1 Illustration of Equation (3.4) . . . 75
3.2 Categorical benchmark datasets. . . . 82
3.3 category weights obtained with *CFE* with $\Phi = 0.05$ on the simulated data for $p = 0.8$. . . . 92
3.4 category weights obtained with *CFE* with $\Phi = 0.05$ on the simulated data for $p = 0.5$. . . . 93
3.6 $p_{yes} = p_{no} = 0.5$ . . . 95
3.5 Simulated dataset to compare the attribute categories frequencies and their weights from *CFE*. . . . 95
3.7 Partition matrix format from Table 3.6. . . . 96

4.1 Compared partitions based on the consistency with the true classes. . . . 107

5.1 Binary variables from the HRS dataset used in our cluster analysis. . . . 129

5.2 Profiles of individuals in each cognitive category based on the attribute categories frequencies. 130
5.3 Mean number of years of education and age per cognitive category. 130
5.4 Mean scores of *CFE* and *cat-ECM* on HRS data set. The values in brackets correspond to the standard deviations. For external measures the scores are computed between the cognition variable (Cogn) and the obtained clusters. 133

A.1 RI scores obtained with *CFE* with different values of β on the datasets. 148
A.2 FRI scores obtained with *CFE* with different values of β on the datasets. 149
A.3 FS scores obtained with *CFE* with different values of β on the datasets. 149
A.4 PC scores obtained with *CFE* with different values of β on the datasets. 150
A.5 PE scores obtained with *CFE* with different values of β on the datasets. 150
A.6 FRI scores obtained with *CFE*, *FC**, and *FCM* for $\beta = 1.9$ on the datasets. This table correspond to the data for the critical difference diagram in Figure 3.5. 151
A.7 RI scores obtained with *CFE*, *FC**, and *FCM* for $\beta = 2$ on the datasets. This table correspond to the data for the critical difference diagram in Figure 3.6. 151

B.1 RI scores obtained with *cat-ECM* with different values of β on the datasets. 154
B.2 FRI scores obtained with *cat-ECM* with different values of β on the datasets. 155
B.3 FS scores obtained with *cat-ECM* with different values of β on the datasets. 155
B.4 PC scores obtained with *cat-ECM* with different values of β on the datasets. 156
B.5 PE scores obtained with *cat-ECM* with different values of β on the datasets. 156

C.1 Frequencies in % of lifestyle factors categories. 161

List of Algorithms

1 *k-means* algorithm (KM) 54
2 *Fuzzy c-means* algorithm (FCM) 55
3 *Evidential c-means* algorithm (ECM) 58
4 *k-modes* algorithm 63
5 *Fuzzy k-modes* algorithm (FKM) 64
6 *Fuzzy center* algorithm (FC) 67
7 *Hard centers updates of FC* algorithm (FC*) 75
8 Categorical fuzzy entropy c-means algorithm (CFE). 79
9 Categorical evidential c-means algorithm (cat-ECM) 105

List of algorithms complexity

Algorithms	Time complexity	Memory complexity
KM	$\mathcal{O}(npcT)$	$\mathcal{O}(np + nc + cp)$
FCM	$\mathcal{O}(ncpT)$	$\mathcal{O}(np + nc + cp)$
ECM	$\mathcal{O}(p^2 n\varrho T + \mathcal{Q}T)$	$\mathcal{O}(np + n\varrho + \varrho p)$
k-modes	$\mathcal{O}(npcT)$	$\mathcal{O}(np + nc + cp)$
FKM	$\mathcal{O}(ncpT)$	$\mathcal{O}(np + nc + cp)$
FC	$\mathcal{O}(nc\mathcal{J}T)$	$\mathcal{O}(np + nc + c\mathcal{J})$
*FC**	$\mathcal{O}(nc\mathcal{J}T)$	$\mathcal{O}(np + nc + c\mathcal{J})$
CFE	$\mathcal{O}(nc\mathcal{J}T)$	$\mathcal{O}(np + nc + c\mathcal{J})$
cat-ECM	$\mathcal{O}(c^2 n\mathcal{J}T + c\varrho\mathcal{J}T + n\varrho\mathcal{J}T)$	$\mathcal{O}(np + n\varrho + \varrho\mathcal{J})$

Introduction

Pattern recognition methods aim at automatically finding patterns and regularities in data. The methods are designed according to the type of data such as matrices, sequences, time series, images, texts, and so on. The input data of these methods may incorporate uncertainty due to the lack of information, conflicting evidence, ambiguity, measurement errors, and belief [2]. In addition to uncertainty, imprecision, vagueness, and inconsistency can be associated with data [3]. Despite their meaning differences, confusion can be made between the preceding four terms. We define below the lexical definitions of these terms according to the Cambridge English Dictionary [1].

According to this dictionary, the term uncertain is defined as *not being able to decide about something, something not known or fixed or not completely unknown.* For example, due to the COVID19 situation, I am uncertain whether to invite people to physically assist in my thesis defense in September 2021. The term imprecise is defined by the dictionary as something *not accurate or exact.* For example, an imprecise statement can be: I am expecting between ten to twenty people to assist in my thesis defense. The term vague is defined as *something not clearly expressed, known, described, or decided.* For example, in the statement "some people will assist in my thesis defense", the type of assistance is vague as it could be physically or virtually. Finally, the term inconsistent is defined as *not agreeing on opposed elements and or something that does not match.* As an example, the statement "I will not recommend people that will physically attend my thesis defense to wear masks and respect the social distancing". This statement is inconsistent with the COVID19 context in 2021[2]. It should be noted that a statement can be imprecise and uncertain at the same time. For instance the statement "It is said that the COVID19 pandemic will end by September 2021", the statement is imprecise by the fact that the information is not precise and uncertain if we don't trust it. For further discussions, we refer the readers to (non-exhaustive references) [4, 5, 6].

[1] https://dictionary.cambridge.org/dictionary/english

[2] At the thesis writing and defense periods, mask-wearing is mandatory in physical meetings.

In this thesis, we propose two new clustering algorithms for categorical data based on different mathematical frameworks for modeling and reasoning with imprecision, vagueness, and uncertainty. The first framework, corresponding to the fuzzy sets theory introduced by Zadeh in [7], is an extension of the hard sets theory that allows gradual assignment of objects in sets instead of binary such as in hard sets. This framework, when applied to clustering, can capture imprecision and vagueness inherent in the input data. The second framework is an extension of the fuzzy sets theory and corresponds to the Dempster-Shafer theory of evidence introduced by Dempster in [8] and developed by Shafer in [9]. The theory is based on belief functions for uncertainty modeling. In addition to the latter theories, the probability theory provides another way of modeling uncertainty. As an application of this theory, Shannon's entropy introduced in [10] has found several applications such as in clustering.

The motivations behind our work are threefold. First, clustering provides a quick way for data exploration and does not require labeled data as in supervised learning. Moreover, the results of cluster analysis due to their interpretability are more accessible to non-experts of the domain. Second, among the tools offering mathematical and computer science frameworks for dealing with uncertain, imprecise, vague, and inconsistent data soft computing has been gaining interest over decades. Particularly, several models based on the fuzzy sets and the Dempster-Shafer theory of evidence have been proposed in the literature and applied to clustering. However, when applied to clustering, many of these methods are designed to fit only one type of data, mostly numerical. Therefore to use these methods for categorical, a conversion that can have several limitations such as the increasing of the dimensions of the data is usually needed. The proposed new methods *categorical fuzzy entropy c-means* and *categorical evidential c-means* respectively referred to as *CFE* and *cat-ECM* are designed to fit categorical data without transformations. While *CFE* uses Shannon's entropy, and the fuzzy sets theory for fuzzy assignments of objects in clusters and the centers of the cluster representations, *cat-ECM* uses the Dempster-Shafer theory of evidence for object labeling. Finally, our work was motivated by the real-world applications of the new methods in developmental sciences to study the influence of lifestyle factors on the cognitive health of older adults.

The organization of the report is as follows: In the first part I, we present the probability, fuzzy sets, and Dempster-Shafer theories in Chapter 1 and a review of clustering algorithms for categorical data in Chapter 2. In the next part II, we present the two new methods *CFE* and *cat-ECM* respectively in Chapter 3 and 4. Finally, in the last part III, we present the applications of the new methods on the analysis of the interactions between lifestyle factors and cognitive health.

Part I

Background

1 — Imprecision and uncertainty theories

Contents

1.1 Hard sets theory as the foundation of imprecision and uncertainty theories 22
1.1.1 Hard sets 22
1.1.2 Properties of hard sets 23
1.1.3 Operations on hard sets 24
1.1.4 Applications of hard sets theory 24
1.1.5 Limitations of hard sets theory 25
1.2 Probability theory 25
1.2.1 Shannon's entropy 26
1.3.1 Fuzzy sets 29
1.3.2 Properties of fuzzy sets 30
1.3.3 Operations on fuzzy sets 30
1.3.4 Applications of fuzzy sets theory 30
1.3 Fuzzy sets theory 29
1.3.5 Limitations of the fuzzy sets theory 31
1.4.1 Representation of evidence 32
1.4.2 Properties of belief functions 38
1.4.3 Operations on belief functions 38
1.4.4 Decision making 39
1.4.5 Uncertainty measures 40
1.4.6 Applications of evidence theory 41
1.4.7 Limitations of the evidence theory 41
1.4 Theory of evidence 32

Introduction

To measure imprecision, vagueness, and uncertainty of information, several theories have been proposed in the literature. Among the main theories, we can cite the fuzzy sets theory introduced by Zadeh [7] that uses fuzzy sets as extensions of hard sets for modeling imprecision and vagueness. For the uncertainty measurement, until the 60s the probability theory and statistics were considered as the only framework for reasoning and modeling uncertainty. Among the probability measures, Shannon's entropy [10] has been used in information theory to measure the uncertainty inherent in a random variable's possible outcomes. In 1978, Zadeh introduced the possibility theory as an extension of the fuzzy sets and fuzzy logic theories to deal with uncertainty [11]. For the same purpose, imprecise probability [12], and the evidence theories [9] have been proposed.

This chapter aims to present the hard sets theory and the mathematical formulation of the probability, fuzzy sets, and evidence theories. Along with the definitions, some properties and applications of the theories are illustrated. A focus on the applications of these theories in clustering is also given.

Remark 1. Along with this report, we consider fuzzy sets in the framework of imprecision and vagueness modeling. Therefore, we dissociate applications of the fuzzy sets theory to other theories such as the possibility theory for uncertainty modeling.

1.1 Hard sets theory as the foundation of imprecision and uncertainty theories

While the fuzzy sets and evidence theories are extensions of classical sets also called hard or crisp sets, the probability theory uses the properties and operations of hard sets to define the probability of collections of elements. In this section, we define the notion of hard sets, some of their properties and operations.

1.1.1 Hard sets

Let $\mathbf{X}$ be a space of points (objects), with a generic element of $\mathbf{X}$ denoted by $\mathbf{x}$ ($\mathbf{X} = \{\mathbf{x}\}$).

Definition 1.1.1 (Hard set)**.** A (hard) set $\mathbf{A}$ is a collection of distinct elements called members. The membership of an object to $\mathbf{A}$ can be

characterized by a function $\mu_{\mathbf{A}}$ such that: $\mathbf{X} \rightarrow \{0, 1\}$

$$\mu_{\mathbf{A}}(\mathbf{x}) = \begin{cases} 1 & \text{if} \quad \mathbf{x} \in \mathbf{A}, \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

Among the properties of the hard sets, we can cite the commutativity, the associativity, the transitivity, and the De Morgan laws of the union and intersection. These properties are presented in the next subsection.

1.1.2 Properties of hard sets

Let **A**, **B**, and **C** be three hard sets.

- The union and the intersection of **A** and **B** are commutative. In other words, the union (respectively the intersection) of **A** and **B** is equal to the union (respectively the intersection) of **B** and **A**.

$$\mathbf{A} \cup \mathbf{B} = \mathbf{B} \cup \mathbf{A}$$

$$\mathbf{A} \cap \mathbf{B} = \mathbf{B} \cap \mathbf{A}.$$

- The union and intersection of hard sets are associative, meaning that the union (respectively the intersection) of hard sets can be decomposed to the union (respectively, the intersection) of tuples of the sets.

$$(\mathbf{A} \cup \mathbf{B}) \cup \mathbf{C} = \mathbf{A} \cup (\mathbf{B} \cup \mathbf{C})$$

$$(\mathbf{A} \cap \mathbf{B}) \cap \mathbf{C} = \mathbf{A} \cap (\mathbf{B} \cap \mathbf{C}).$$

- The inclusion of hard sets is transitive, i.e., if a set A is contained in another set B and itself contained in a set C, then A is contained in C.

$$\text{If } \mathbf{A} \subseteq \mathbf{B} \subseteq \mathbf{C}, \text{ then } \mathbf{A} \subseteq \mathbf{C}.$$

- De Morgan's law of hard sets are described as follows: the complement of the union (respectively the intersection) of two hard sets is equal to the intersection (respectively the union) of the complements of the sets.

$$\neg(\mathbf{A} \cup \mathbf{B}) = \neg\mathbf{B} \cap \neg\mathbf{A}$$

$$\neg(\mathbf{A} \cap \mathbf{B}) = \neg\mathbf{B} \cup \neg\mathbf{A}.$$

In the next subsection, we present the basics operations of hard sets.

1.1.3 Operations on hard sets

Operations on sets help to create new sets from existing ones. The main operations are equality, containment, intersection, union, and complement. They are respectively defined by Equations (1.2a), (1.2b), (1.2c), (1.2d), and (1.2e).

Let $\mathbf{A}$ and $\mathbf{B}$ be two hard sets and $\mathbf{x} \in \mathbf{X}$,

$$\text{Equality}: \ \mathbf{A} = \mathbf{B} \Leftrightarrow \mathbf{A} \subseteq \mathbf{B} \text{ and } \mathbf{B} \subseteq \mathbf{A}. \tag{1.2a}$$
$$\text{Containment}: \ \mathbf{A} \subset \mathbf{B} \Leftrightarrow \mathbf{x} \in \mathbf{A} \Rightarrow \mathbf{x} \in \mathbf{B}. \tag{1.2b}$$
$$\text{Intersection}: \ \mathbf{A} \cap \mathbf{B} = \{\mathbf{x} \in \mathbf{X} | \mathbf{x} \in \mathbf{A} \text{ and } \mathbf{x} \in \mathbf{B}\}. \tag{1.2c}$$
$$\text{Union}: \ \mathbf{A} \cup \mathbf{B} = \{\mathbf{x} \in \mathbf{X} | \mathbf{x} \in \mathbf{A} \text{ or } \mathbf{x} \in \mathbf{B}\}. \tag{1.2d}$$
$$\text{Complement}: \ \neg\mathbf{A} = \{\mathbf{x} \in \mathbf{X} | \mathbf{x} \notin \mathbf{A}\}. \tag{1.2e}$$

Equation (1.2a) expresses the equality of two hard sets: two hard sets $\mathbf{A}$ and $\mathbf{B}$ are equal if and only if $\mathbf{A}$ is included in $\mathbf{B}$ and $\mathbf{B}$ is included in $\mathbf{A}$. The containment expressed by $\mathbf{A}$ is contained in $\mathbf{B}$ (1.2b) occurs if and only if any element of $\mathbf{A}$ is included in $\mathbf{B}$. The intersection of two sets $\mathbf{A}$ and $\mathbf{B}$ (1.2c) corresponds to the elements that are both in $\mathbf{A}$ and $\mathbf{B}$ whereas their union (1.2d) corresponds to a new set that contains all the elements of $\mathbf{A}$ and $\mathbf{B}$. Finally, the complementary of a set $\mathbf{A}$ (1.2e) denoted here by $\neg\mathbf{A}$ contains all the elements not in $\mathbf{A}$.

A dual representation of Equation (1.2) can be defined with the membership function μ as follows:

$$\text{Equality}: \ \mu_{\mathbf{A}} = \mu_{\mathbf{B}} \quad \forall \mathbf{x} \in \mathbf{X} \quad \mu_{\mathbf{A}}(\mathbf{x}) = \mu_{\mathbf{B}}(\mathbf{x}). \tag{1.3a}$$
$$\text{Containment}: \ \mu_{\mathbf{A}} \leq \mu_{\mathbf{B}} \quad \forall \mathbf{x} \in \mathbf{X} : \mu_{\mathbf{A}}(\mathbf{x}) \leq \mu_{\mathbf{B}}(\mathbf{x}). \tag{1.3b}$$
$$\text{Intersection}: \ \mu_{\mathbf{A}\cap\mathbf{B}}(\mathbf{x}) = \min[\mu_{\mathbf{A}}(x), \ \mu_{\mathbf{B}}(\mathbf{x})], \quad \mathbf{x} \in \mathbf{X}. \tag{1.3c}$$
$$\text{Union}: \ \mu_{\mathbf{A}\cup\mathbf{B}(\mathbf{x})} = \max[\mu_{\mathbf{A}}(\mathbf{x}), \ \mu_{\mathbf{B}}(\mathbf{x})], \quad \mathbf{x} \in \mathbf{X}. \tag{1.3d}$$
$$\text{Complement}: \ \mu_{\neg\mathbf{A}}(\mathbf{x}) = 1 - \mu_{\mathbf{A}}(\mathbf{x}), \quad \mathbf{x} \in \mathbf{X}. \tag{1.3e}$$

Operations on hard sets can be applied in different domains such as in clustering. In the following subsection, we present a brief overview of the applications of hard sets to this domain.

1.1.4 Applications of hard sets theory

The hard sets theory applications are common in mathematics particularly in the construction of relations between mathematical objects. The theory also serves as the base of the probability theory. In clustering, the hard sets theory can be used to generate a hard c-partition (see in Section 2.1.1) from an input dataset with c being an integer greater than 2. The most popular

partitioning clustering method for generating such partitions is the *k-means* algorithm presented in Section 2.2.1. Due to the limitations of this method in handling only numerical data, it has been extended to categorical data. We present some of the extensions in Chapter 2.

The binary representation of hard sets membership degrees limits applications of the hard sets theory on problems where the solutions can have fuzzy boundaries. In the following subsection, we illustrate this limitation.

1.1.5 Limitations of hard sets theory

Let's consider the statement "The COVID19 pandemic is ending before December 2021!"[1], the ending month is not precise, and all the months from January to November 2021 are possible. With the hard sets theory, the membership function will indicate for each month whether the pandemic is ending 1 or not 0. However, this representation of the information does not reflect reality because we do not know precisely the exact month at which the pandemic is ending. There is therefore a need for new theories that can model imprecise and vague information (data). In other words, for instance, a theory that will assign a gradual degree for instance varying from 0 to 1 to indicate the possible ending month of the pandemic. In the literature, many theories have been proposed for that purpose. Among them, we have the fuzzy sets and evidence theories discussed respectively in Section 1.3 and 1.4.

Before presenting the fuzzy sets and evidence theories, we first review the probability theory and Shannon entropy which can be used to measure uncertainty.

1.2 Probability theory

Until the emergence of new theories in the 50s, the probability theory was considered as the only framework for modeling and reasoning with uncertainty. As the foundation of statistics, the probability theory is still the most accepted theory in the literature for uncertainty modeling [2]. In this section, we present a brief review of this theory and its application to information theory.

The probability theory measures the uncertainty of the outcome of a random event through a mapping function from the set of possible outcomes to the interval $[0, 1]$. Depending on the type of outcomes which can be discrete or continuous, the probability distribution is said respectively discrete or continuous. For simplicity, we limit the review to the discrete case.

[1] By considering we are in January 2021.

Let's define a probability function.

Definition 1.2.1 (Probability function)**.** Let Ω called sample space be a finite (hard) set of all possible outcomes of a random experiment and $\mathbf{A}$ and $\mathbf{B}$ be two subsets of Ω. A probability function is a function $P : \Omega \longrightarrow [0, 1]$ such that

$$\begin{cases} P(\emptyset) = 0, \\ P(\Omega) = 1, \\ P(\mathbf{A} \cup \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \cap \mathbf{B}). \end{cases} \tag{1.4}$$

■

To quantify uncertainty in the probability theory framework several measures have been proposed in the literature. Among these measures, entropy quantifies the disorder of the state of a system. The former entropy measure initially introduced by Shannon in [10] as a measure of information has been adapted to the probability theory. In the next subsection, we present this measure in the framework of the probability theory.

1.2.1 Shannon's entropy

In information theory (i.e., mathematical theory of communication) Shannon's entropy is a measure for quantifying the amount of uncertainty associated with the outcome of a random variable. It was introduced by Shannon in 1948 [10].

Definition 1.2.2 (Shannon's entropy)**.** Let $P = (p_1, \ldots, p_n)$ be a finite probability distribution of a random variable $Y = (y_1, \ldots, y_n)$ such that $\sum_{k=1}^{n} p_k = 1$. The Shannon's entropy associated to P usually denoted by $H_n(p_1, \ldots, p_n)$ is defined by

$$H_n(p_1, \ldots, p_n) = -\sum_{k=1}^{n} p_k \log_b(p_k), \tag{1.5}$$

where b is the base of the logarithm. In this report, we consider the natural logarithm, i.e., $b = e$. It will be referred as ln.

In the following example, we consider the use of Shannon's entropy to measure the uncertainty of the effectiveness of COVID19 vaccines.

Example 1.2.1. Let Y = (Pfizer, AstraZeneca) be a set entitled "the two most used COVID19 vaccines in France", and $p_1 = 0.8$ and $p_2 = 0.2$ be respectively the probability of success of the vaccines on UK's COVID19 variant. Shannon's entropy $H_2(p_1, p_2)$ representing the uncertainty of the success of the two vaccines on the variant is given by

$$\begin{aligned} H_2(p_1, p_2) = & -0.8\ln(0.8) - 0.2\ln(0.2) \\ = & 0.5. \end{aligned}$$

In the framework of the probability theory, Shannon's entropy is described by different properties. In the next subsection, we present some of them.

Properties

Shannon's entropy defined in Equation (1.5) possesses several properties. Among them, we have the non-negativity (1), the symmetry (1.6), the monotonicity (6), the additivity (3), and the concavity (4).

1. $H_n(P) \geq 0$, with equality when $\exists k$ such that $p_k = 1$. A zero entropy implies that the process is deterministic.

2. The entropy $H_n(p_1, \ldots, p_n)$ does not depend on the order of p_i's and under an arbitrary permutation $\{\alpha_1, \ldots, \alpha_n\}$ of the set $\{1, \ldots, n\}$

$$H_n(p_1, \ldots, p_n) = H_n(p_{\alpha_1}, \ldots, p_{\alpha_n}). \tag{1.6}$$

3. Consider two probability distributions $P = (p_1, \ldots, p_n)$ and $E = (e_1, \ldots, e_n)$ associated with independent random variables Z and Y. The joint probability distribution of Z and Y is given by

$$\mathrm{P}(Z = z_i, Y = y_j) = p_i e_j, \forall i \in \{1, \ldots, n\}, \forall j \in \{1, \ldots, m\}. \tag{1.7}$$

   From (1.7), the entropy of the joint distribution $\mathcal{H}_{nm} = H_{nm}(p_1e_1; \ldots; p_1e_m; \ldots; p_ne_1; \ldots; p_ne_m)$ corresponds to the sum of respective entropies associated with the independent random variables:

$$\mathcal{H}_{nm} = H_n(p_1, \ldots, p_n) + H_m(e_1, \ldots, e_n). \tag{1.8}$$

4. H_n is a concave function of p_i's therefore its local maximum corresponds to the global maximum.

5. $H_n(P) \leq H_n(\frac{1}{n}, \ldots, \frac{1}{n})$. It implies that maximum entropy or maximum disorder is reached when all probabilities are equal. For instance the maximum entropy of the weights associated to the attributes categories in the fuzzy centers algorithm described in Section 2.2.8 is reached when all the weights are equal to $\frac{1}{n_l}$.

6. H_n is a monotonically decreasing function of p_i's (i.e., if the probabilities p_i increase, H_n decreases).

We refer the readers to [13, 14] for detailed discussions on the properties of Shannon's entropy.

Due to its mathematical properties and its simplicity in calculus, Shannon's entropy has been used in a wide range of domains. In the next subsection, we present some of these applications.

Applications

Although initially proposed in the context of communication theory, Shannon's entropy has been applied in a wide range of problems ranging from physical [15], engineering [16], finance [17], biological [18], social [19] and environmental sciences [20] to pattern recognition [21]. In cluster analysis, Shannon's entropy has been used in two main applications: as an internal measure and as a penalization function. The former application uses entropy to quantify the goodness of fit of a clustering algorithm whereas the latter application uses the entropy to regularize data or a learning process. In Section 2.2.4, we present some examples of applications of Shannon's entropy in clustering as a penalization function.

As Shannon's entropy, fuzzy sets have been used in several applications. In the next section, we provide a brief description of the theory and its applications.

1.3 Fuzzy sets theory

Due to the limitations of the hard sets theory to handle fuzzy boundaries, a new theory called fuzzy sets was introduced as an extension of hard sets theory for modeling imprecise and vague data. In the next subsections, we present the mathematical definition of fuzzy sets, some of their properties, operations, and applications.

1.3.1 Fuzzy sets

A piece of information is said to be fuzzy or vague when its boundary is not clear. Contrary to the hard sets theory in which elements memberships are binary, the fuzzy sets allow gradual assignment of these elements. The theory was introduced independently by Zadeh [7] and Klaua [22] in 1965. The original idea of the theory is to model imprecise information expressed in natural language. Zadeh referred to this in [7] by "more often than not, the classes of objects encountered in the real physical world do not have precisely defined criteria of membership". To illustrate the ambiguity in natural language, we can consider the term "young" in the context of aging. Some people may define a young person as someone whose age is below 25 years old, whereas for other people, this threshold may not correspond to their perception of the term.

Definition 1.3.1 (Fuzzy set)**.** A fuzzy set $\mathcal{A}$ in $\mathbf{X}$ is characterized by a membership (characteristic) function $\mu_{\mathcal{A}}$ which associates to each element in $\mathbf{X}$ a real number in the interval [0, 1], with the value of $\mu_{\mathcal{A}}$ at $\mathbf{x}$ representing the membership degree of $\mathbf{x}$ in $\mathcal{A}$. The fuzzy set $\mathcal{A}$ is usually represented by the pair $(\mathbf{X}, \mu_{\mathcal{A}})$. When X is finite (i.e., $\mathbf{X} = \{x_1, \ldots, \mathbf{x}_n\}$), $\mathcal{A}$ is often denoted by:

$$\mathcal{A} = \{\mu_{\mathcal{A}}(\mathbf{x_1})/\mathbf{x_1}, \ldots, \mu_{\mathcal{A}}(\mathbf{x_n})/\mathbf{x_n}\}. \tag{1.9}$$

■

Example 1.3.1 (Fuzzy set)**.** Considering the preceding statement "The COVID19 pandemic is ending by December 2021!", this statement is imprecise as the ending month of the pandemic is not clear. To model this imprecision, a fuzzy set can be used by associating membership values indicating the possible ending of the pandemic to each month before December such as: **0.02** to January (a_1), **0.02** to February (a_2), ..., **0.5** to October (a_{10}), **0.7** to November (a_{11}).

Let $\mathcal{A}$ be the fuzzy set associated to this event, following the notation in Equation (1.9), we have:

$$\mathcal{A} = \{\mathbf{0.02}/a_1, \mathbf{0.02}/a_2, \ldots, \mathbf{0.5}/a_{10}, \mathbf{0.7}/a_{11}\}.$$

In the following subsection, we describe the main properties of fuzzy sets.

1.3.2 Properties of fuzzy sets

As an extension of the hard sets theory, the fuzzy sets theory inherits all the properties of hard sets. Hence the properties of hard sets described in Section 1.1 also apply to fuzzy sets.

In the next subsection, basics operations on fuzzy sets are discussed.

1.3.3 Operations on fuzzy sets

The main operations of fuzzy sets are defined in the same way as hard sets in Equation (1.3). Apart from the equality operation, the results of the remaining operations are new fuzzy sets. Furthermore, the minimum and maximum functions in the Equations (1.3c) and (1.3d) can be respectively generalized with the t-norm and s-norm (i.e., t-conorm) [23] for fuzzy sets.

Depending on the value of the membership degree function of fuzzy sets, crisp sets can be related to the fuzzy sets through such as the notion of support (Supp) and kernel (Kern) defined respectively by Equation (1.10) and (1.11).

$$Supp(\mathcal{A}) = \{x \in \mathbf{X} |\ \mu(x) > 0\}. \tag{1.10}$$

$$Kern(\mathcal{A}) = \{x \in \mathbf{X} |\ \mu(x) = 1\}. \tag{1.11}$$

With their ability to provide solutions to problems with fuzzy boundaries, the fuzzy sets theory has successful applications in a wide range of domains. In the next subsection, we present some of these applications with a focus on fuzzy clustering.

1.3.4 Applications of fuzzy sets theory

The fuzzy sets theory has been applied both to other formal theories and real problems. As examples of applications, we have fuzzy logic and approximate reasoning, expert systems, control, databases and queries, data analysis, engineering, and management [2]. In the latter book, Zimmermann classifies the applications of the fuzzy sets theory into four families as follows:

- Applications to mathematics, that is, generalizations of traditional mathematics such as topology, graph theory, algebra, logic, and so on.
- Applications to algorithms such as clustering methods, control algorithms, mathematical programming, and so on.
- Applications to standard models such as "the transportation model", "inventory control models", "maintenance models", and so on.
- Applications to real-world problems of different kinds such as engineering and management.

We refer the readers to [13, 24, 2, 25] for detailed reviews on applications of the theory to other theories and real problems.

In clustering, the fuzzy sets theory has been used among others for fuzzy assignments of objects in clusters and cluster center representations. In the former application, fuzzy partitions can be obtained with various clustering methods. The most popular method is the *fuzzy c-means* proposed by Bezdek [26] in the 80s. Several extensions of the *fuzzy c-means* have been proposed in the literature for handling different types of data: categorical, time series, image, relational, and so on. Among the categorical extensions of the *fuzzy c-means*, we have the *fuzzy k-modes* [27] algorithm which extends the *k-modes* [28, 29], the *fuzzy centroids clustering* [30] which uses the fuzzy sets theory for representing the centroids. The preceding algorithms are presented in Chapter 2.

Despite offering flexible ways of tackling fuzzy boundaries problems, the fuzzy sets theory has some limitations discussed in the following paragraph.

1.3.5 Limitations of the fuzzy sets theory

In clustering, depending on the application, fuzzy partition degrees can be hard to interpret. For instance they do not express the belief of object assignment in the clusters. Indeed, in some real-life scenarios, the basics interpretations of fuzzy membership degrees as similarity degrees might not be sufficient. For instance, if the membership degree to a cluster is one, one would like to say that it is certain that the corresponding objects belong to the cluster. Therefore, there is a need for new theories to model the uncertainty of the object's membership degrees. In the literature, the Dempster-Shafer theory of evidence was proposed for that purpose. When applied to clustering, the latter theory provides measures to quantify the uncertainty of object labeling.

In the next section, we briefly review the Dempster-Shafer theory of evidence.

1.4 Theory of evidence

The belief functions theory also called Dempster-Shafer theory or theory of evidence is a mathematical theory introduced by Glenn Shafer in 1976 [9] as a new approach of uncertainty modeling. The theory is described by Shafer as a reinterpretation of Arthur P. Dempster's work [8] in the context of statistical inference on lower and upper bounds of probabilities: "...it offers a reinterpretation of Dempster's work, a reinterpretation that identifies his lower probabilities as epistemic probabilities or degrees of belief takes the rule for combining such degrees of belief as fundamental and abandons the idea that they arise as lower bounds over classes of Bayesian probabilities" [8]. The theory considered as a generalization of the Bayesian theory of subjective probability has been developed and popularized by Smets and Kennes under the name of the transferable belief model [31]. This model is an axiomatic approach of belief functions where a belief function can be held at two levels [31]: (1) a credal level where beliefs are entertained and quantified beliefs functions, (2) a pignistic level where beliefs can be used to make decisions and are quantified by probability functions.

To introduce the notion of belief functions, let's consider again the statement "It is said that the COVID19 pandemic will end before December 2021!" in the general introduction. We saw that this statement is both imprecise due to the imprecise date of ending of the pandemic and uncertain if the source of the information is unreliable. Let's consider that our subjective probability of the reliability of the source is 0.8. Hence the subjective probability of being unreliable is 0.2. Since we are considering probabilities, the two probabilities sum up to 1. The statement is true if the source of the information is reliable but not necessarily false if it is unreliable. With the belief functions theory we say that with the statement alone, we have a 0.8 degree of belief that the pandemic will end by December 2021 and a 0 degree of belief (not a 0.2 degree of belief) that the pandemic will not end by December 2021. Contrary to the probability theory, the 0 here does not mean that we are not sure that the pandemic will not end by December 2021 but we have no reason to believe what is advanced in our example. The previous example is inspired by Shafer's in [32].

1.4.1 Representation of evidence

Let $\Omega = \{\omega_1, \ldots, \omega_k, \ldots, \omega_c\}$ be a finite set of the states of a system called the frame of discernment. In clustering, for instance, Ω represents the set of desired clusters and ω_k the subsets of clusters. Partial knowledge about these states can be characterized by mass function also called basic belief assignment defined as follows:

Definition 1.4.1 (Mass function)**.** A mass (m) can be associated to any subset A of Ω such that $m : 2^{\Omega} \rightarrow [0, 1]$ and

$$\sum_{\mathbf{A} \subseteq \Omega} m(\mathbf{A}) = 1. \tag{1.12}$$

■

Depending on the values of m called basic belief masses special mass functions can be derived. For instance, a subset $\mathbf{A}$ of Ω is called a focal set when the corresponding mass function is positive ($m(\mathbf{A}) > 0$). Special cases of focal sets are the Bayesian, logical, and vacuous mass functions defined as follows:

- A mass function is said to be Bayesian when all focal sets of Ω are singletons (sets with cardinality one). In this case, the mass function corresponds to a probability distribution.
- A mass function is said to be logical when it has only one focal set.
- A mass function is said to be vacuous if $m(\Omega) = 1$. In this case, the mass function represents a completely uninformative piece of evidence (i.e., complete ignorance).

Example 1.4.1 (Mass function)**.** Let Ω be the set of UK's, South Africa and Brazil COVID19 variants and the original virus: $\Omega = \{\text{Original}, \text{UK}, \text{South Africa}, \text{Brazil}\}$. Let's suppose that the Pfizer vaccine is efficient with a probability of 0.9 on the original virus and the UK's variant. We know that the vaccine is efficient with a probability of at least 10% with any kind of variants including the original virus. We can model this information using a mass function m as follows:

$$m(\{\text{Original}, \text{UK}\}) = 0.9$$
$$m(\Omega) = 0.1.$$

The set $\{\text{Original}, \text{UK}\}$ here is a focal set and $m(\Omega)$ represents the degree of ignorance allocated to effectiveness of the vaccine on the variants.

From Equation (1.12), the value of the mass allocated to the empty set can be positive. In the original work of Shafer in [9], this value is constrained to be 0 and represents a closed-world assumption, i.e., the frame of

discernment is exhaustive. In contrast, in the open-world assumption [33], the value of $m(\emptyset) > 0$ is interpreted as the degree of the state of the system not underlying in the frame of discernment.

In general, the closed-world assumption is considered and when $m(\emptyset) > 0$, different normalization procedures such as Dempster's or Yager's rules can be performed.

Let $\mathbf{A}$ be a subset of Ω such that $A \neq \emptyset$, $m(\mathbf{A})$ the mass function of $\mathbf{A}$ and $m^*(A)$ the normalized mass function of $\mathbf{A}$. The Dempster's normalization procedure of m given by the Equation (1.13) consists in allocating the mass of the empty set to $\mathbf{A}$:

$$m^*(A) = \begin{cases} \dfrac{m(\mathbf{A})}{1 - m(\emptyset)} & \text{if} \quad A \neq \emptyset \\ 0 \quad \text{otherwise.} \end{cases} \tag{1.13}$$

whereas Yager's normalization procedure given by the Equation (1.14) consists in assigning the mass of 0 to $m^*(A)$ when $\mathbf{A}$ is the empty set, the mass of Ω and the empty set when A corresponds to Ω and $m(\mathbf{A})$ in the other cases.

$$m^*(A) = \begin{cases} 0 \quad \text{if} \quad \mathbf{A} = \emptyset \\ m(\Omega) + m(\emptyset) \quad \text{if} \quad A = \Omega \\ m(\mathbf{A}) \quad \text{otherwise.} \end{cases} \tag{1.14}$$

In cluster analysis, a normalization of a partition obtained with the Dempster-Shafer theory can be performed when transforming the later partition to other types of partitions such as fuzzy (see Chapter 2).

In some scenarios, the evidence of a subset of Ω can be used to determine if the latter implies or contradicts the evidence of another subset of Ω. This can be achieved by using belief and plausibility functions defined as follows.

Let $\mathbf{A}$ and $\mathbf{B}$ be two subsets of Ω such that $\mathbf{A} \subseteq \mathbf{B}$, if the evidence is true for $\mathbf{A}$ then the evidence is said to support $\mathbf{B}$. In other words, the evidence of a set can be characterized by the evidence of all its nonempty subsets. Similarly, if the evidence does not support the complement of $\mathbf{B}$ referred to as $\neg B$, the evidence is said to be consistent with $\mathbf{B}$.

Definition 1.4.2 (Belief function)**.** Let m be a mass function, the belief in $\mathbf{B}$ (i.e., the probability that the evidence implies B) is given by the function $bel : 2^{\Omega} \to [0, 1]$

$$bel(B) = \sum_{\mathbf{A} \subseteq \mathbf{B}, \mathbf{A} \neq \emptyset} m(\mathbf{A}) \quad \text{and} \quad bel(\emptyset) = 0. \tag{1.15}$$

■

Definition 1.4.3 (Plausibility function)**.** Let m be a mass function, the plausibility in a set $\mathbf{B}$ (i.e., the probability that the evidence does not contradict $\mathbf{B}$) is given by the function $pl : 2^{\Omega} \rightarrow [0, 1]$

$$pl(\mathbf{B}) = \sum_{\mathbf{A} \cap \mathbf{B} \neq \emptyset} m(\mathbf{A}) = 1 - bel(\neg \mathbf{B}) \quad \text{and} \quad pl(\emptyset) = 0. \tag{1.16}$$

■

From the **Definition** 1.4.2 and 1.4.3, it can be noted that the belief and plausibility degrees are equivalent representations of a mass function. Furthermore, the two functions respectively represent the lower and upper degree of belief of uncertainty. For any focal element $A \subseteq \Omega$, the interval $[bel(A), pl(A)]$ is called a belief interval.

In the following example, we provide a simulated application of the belief and plausibility functions to model the uncertainty of the effectiveness of COVID19 vaccines.

Example 1.4.2 (Belief and plausibility)**.** The European Union (EU) would like to conduct studies on the effectiveness before allowing their usage of Pfizer, Moderna, AstraZeneca (random sample of vaccines) vaccines on the original, UK's, South Africa's and Brazil's variants of the COVID. A prior study was conducted a few months ago on the effectiveness of Pfizer, Moderna, and AstraZeneca vaccines on the original COVID19 virus. From this study, the EU knows that these vaccines have respectively a probability of failure of 0.1, 0.2 for the Pfizer and Moderna vaccines on the original virus. From the same study, it is known that having a good combination of vaccines can reduce the probability of failure. For instance, when the Pfizer's and Moderna vaccines have combined the probability of failure is 0.18 and when the Pfizer's, and AstraZeneca's are combined the probability of failure is 0.15 when the Moderna and AstraZeneca are combined the probability is 0.2 and when all the vaccines are combined, the probability of failure is 0.07. As AstraZeneca is the latest developed vaccine, the EU wants to take some caution in including this vaccine in the study. The laboratory that developed the vaccine assure that their vaccine is very efficient and has a probability of failure of only 0.1 on the original virus. To avoid spending

more time verifying the information given by the laboratory, the EU decided to trust it. Two months later, the EU obtained the results of the failure of the vaccines on the variants. The probabilities are reported in Table 1.1.

Vaccines	Original	UK	South Africa	Brazil	Mean
$\{\emptyset\}$	0	0	0.19	0	0
{Pfizer}	0.1	0.2	0.3	0.05	0.1625
{Moderna}	0.2	0.2	0.1	0.45	0.2375
{AstraZeneca}	0.1	0.1	0.2	0.15	0.1375
{Pfizer, Moderna}	0.18	0.18	0.1	0.15	0.1525
{Pfizer, AstraZeneca}	0.15	0.02	0.01	0.05	0.0575
{Moderna, AstraZeneca}	0.2	0.2	0.08	0.1	0.1450
{Pfizer, Moderna, AstraZeneca}	0.07	0.1	0.02	0.05	0.0600

Table 1.1: Probability of failure of the vaccines on all the studied virus of COVID.

The EU would like now to know the most efficient vaccine or combination of vaccines (at most 2 vaccines) on both the original virus and the studied variants. For precautions of using the obtained efficient vaccine(s), the EU would like to also have a confidence interval.

To answer the EU's requests, belief and plausibility degrees over all the vaccines and combinations of vaccines can be computed to determine the vaccine or combination that for instance maximizes the overall degrees. These degrees will then serve as the lower and upper bounds of the probability of effectiveness of the vaccines.

Let's consider Ω the set of the following COVID19 vaccines: $\Omega = \{\text{Pfizer}, \text{Moderna}, \text{AstraZeneca}\}$. We have

$$2^{\Omega} = \{\emptyset, \{\text{Pfizer}\}, \{\text{Moderna}\}, \Omega\}.$$

The computation the belief and plausibility of {Pfizer} are described as follows:

1. Belief
   From Equation (1.15) we have:

$$bel_{\text{Original}}(\{\text{Pfizer}\}) = m_{\text{Original}}(\{\text{Pfizer}\}) = 0.1.$$

2. Plausibility

From Equation (1.16) we have:

$$\begin{aligned} pl_{\text{Original}}(\{\text{Pfizer}\}) &= m_{\text{Original}}(\{\text{Pfizer}\}) \\ &\quad + m_{\text{Original}}(\{\text{Pfizer, Moderna}\}) \\ &\quad + m_{\text{Original}}(\{\text{Pfizer, AstraZeneca}\}) \\ &\quad + m_{\text{Original}}(\{\text{Pfizer, , Moderna, AstraZeneca}\}) \\ &= 0.1 + 0.18 + 0.15 + 0.07 \\ &= 0.5. \end{aligned}$$

As the belief and plausibility degrees correspond to the lower and upper bounds of the probability of failure, if $p_{\text{Original}}(\{\text{Pfizer}\})$ is the probability of failure of the Pfizer's vaccine on the original virus, we then have from 1 and 2 $p_{\text{Original}}(\{\text{Pfizer}\}) \in [0.1, 0.5]$.

By computing of the belief and plausibility degrees of all the subsets of vaccines the tables 1.2 and 1.3 are obtained.

Vaccines	bel_{Original}	bel_{UK}	$bel_{\text{South Africa}}$	bel_{Brazil}	bel_{mean}
$\{\emptyset\}$	0	0	0	0	0
{Pfizer}	0.1	0.3	0.3	0.05	0.1875
{Moderna}	0.2	0.1	0.1	0.45	0.2125
{AstraZeneca}	0.1	0.2	0.2	0.15	0.1625
{Pfizer, Moderna}	0.48	0.58	0.5	0.65	0.5525
{Pfizer, AstraZeneca}	0.35	0.32	0.51	0.25	0.3575
{Moderna, AstraZeneca}	0.5	0.5	0.38	0.7	0.52
{Pfizer, Moderna, AstraZeneca}	1	1	1	1	1

Table 1.2: Belief of failure of all vaccines on the variants.

Vaccines	pl_{Original}	pl_{UK}	$pl_{\text{South Africa}}$	pl_{Brazil}	pl_{mean}
$\{\emptyset\}$	0	0	0	0	0
{Pfizer}	0.5	0.5	0.43	0.3	0.4325
{Moderna}	0.65	0.68	0.3	0.45	0.52
{AstraZeneca}	0.52	0.42	0.31	0.75	0.5
{Pfizer, Moderna}	0.9	0.9	0.8	0.85	0.8625
{Pfizer, AstraZeneca}	0.8	0.8	0.9	0.55	0.7625
{Moderna, AstraZeneca}	0.9	0.8	0.7	0.95	0.8375
{Pfizer, Moderna, AstraZeneca}	1	1	0.81	1	1

Table 1.3: Plausibility of failure of all vaccines on the variants.

Based on the hypothesis that the best vaccine corresponds to the vaccine with the lowest degree of failure on overall on the COVID19 variants, from Tables 1.1, 1.2 and 1.3, it can be deduced that, the best vaccine is Pfizer with

a mean of probability of failure of 16.25% and a lower and upper bounds of respectively 0.1875 and 0.4325.

In the following subsection, we present the main properties of belief functions.

1.4.2 Properties of belief functions

Among the main properties of belief functions, monotonicity occupies an important role. Indeed, the function *bel* is a completely monotone capacity. In other words the function *bel* verifies:

$$bel(\emptyset) = 0, \tag{1.17a}$$

$$bel(\Omega) = 1, \tag{1.17b}$$

$$bel\Big(\bigcup_{i=1}^{k} \mathbf{A_i}\Big) \geq \sum_{I \subseteq \{1,\dots,k\}, I \neq \emptyset} (-1)^{|I|+1} bel\Big(\bigcap_{i \in I} \mathbf{A_i}\Big), \tag{1.17c}$$

for any $k \geq 2$ and for any subsets $\mathbf{A_i}, \dots \mathbf{A_k}$ in 2^{Ω}.

As consequence of this property, a unique mass function can be associated to a belief function as follows:

$$m(\mathbf{A}) = \sum_{\mathbf{B} \subseteq \mathbf{A}, \mathbf{B} \neq \emptyset} (-1)^{|\mathbf{A}|-|\mathbf{B}|} bel(\mathbf{B}). \tag{1.18}$$

In addition to being characterized by the preceding properties, several operations can be applied to belief functions such as described in the next subsection.

1.4.3 Operations on belief functions

As for the fuzzy sets theory, operations on belief functions can be used in order to create new belief functions. Among the main operations, the combination or fusion of two pieces of evidence from an independent frame of discernment can be used to have a more informative mass function. For that purpose, Dempster's rule of combination may be employed. Depending on the nature of the information of the sources, the combination can be conjunctive when the sources are reliable and disjunctive when the sources are not reliable.

Definition 1.4.4 (Dempster's rule of combination)**.** Let m_1 and m_2 be two mass functions on the same frame Ω induced by two independent pieces of evidence. The Dempter's rule of combination is defined by

$$(m_1 \oplus m_2)(\mathbf{A}) = \frac{1}{1-K} \sum_{\mathbf{B}\cap\mathbf{C}=\mathbf{A}} m_1(\mathbf{A})m_2(\mathbf{C}) \quad \forall \mathbf{A} \neq \emptyset, \tag{1.19}$$

where

$$K = \sum_{\mathbf{B}\cap\mathbf{C}=\emptyset} m_1(\mathbf{A})m_2(\mathbf{C}) \tag{1.20}$$

is the degree of conflict between m_1 and m_2.

The conjunctive and disjunctive operations can be deduced from Dempster's rule of combination as follows:

Definition 1.4.5 (Conjunctive and disjunctive combinations)**.** Let m_1 and m_2 be two distinct mass functions.

The conjunctive combination of m_1 and m_2 denoted by the $m_1 \cap m_2$ is given by

$$(m_1 \cap m_2)(\mathbf{A}) = \sum_{\mathbf{B}\cap\mathbf{C}=\mathbf{A}} m_1(\mathbf{A})m_2(\mathbf{C}) \quad \forall \mathbf{A} \subseteq \Omega. \tag{1.21}$$

whereas the disjunctive combination of m_1 and m_2 denoted by the $m_1 \cup m_2$ is given by

$$(m_1 \cup m_2)(\mathbf{A}) = \sum_{\mathbf{B}\cup\mathbf{C}=\mathbf{A}} m_1(\mathbf{A})m_2(\mathbf{C}) \quad \forall \mathbf{A} \subseteq \Omega. \tag{1.22}$$

As a generalization of the Bayesian subjective probability theory, the Dempster-Shafer theory of evidence provides transformation procedures for converting belief degrees to probability degrees. An example of this transformation is discussed in the next subsection.

1.4.4 Decision making

In Example 1.4.2 the best efficient COVID19 vaccine from the sample is obtained by taking the vaccine having the lowest plausibility of failure. The method used in the decision making is from [34] which consists of choosing the singleton of Ω with the highest plausibility (in our case the lowest plausibility as we want to minimize the probability of failure). More generally for

decision-making on belief functions, a formal approach would be to transform the belief function models to probability models. For that purpose, Smets in [35] introduced the pignistic transformation in the framework of belief functions. In the transformation function denoted by BetP (for betting probability) and given by Equation (1.23) each mass of belief $m(\mathbf{A})$ is proportionally distributed among the elements of A.

$$BetP(\omega) = \sum_{\omega \in \mathbf{A}} \frac{m(\mathbf{A})}{|\mathbf{A}|} \quad \forall \omega \in \Omega \text{ and } m(\emptyset) = 0, \tag{1.23}$$

where $|\mathbf{A}|$ corresponds to the cardinality of $\mathbf{A} \subseteq \Omega$.

In clustering, this transformation can be used to convert an evidential partition a fuzzy partition (see Chapter 4).

As an uncertainty theory, the evidence theory provides several measures to quantify uncertainty. In the next subsection, some of them are presented.

1.4.5 Uncertainty measures

Shannon's entropy presented in Section 1.2 when applied to a fuzzy partition quantifies the degree of disorder in the corresponding partition. Likewise, to describe the uncertainty in a *bba*, entropy-like measures have been proposed in the framework of belief functions such as the ambiguity [36] and the aggregated uncertainty [37]. In [38], Klir and Wierman proposed a nonspecificity measure given by Equation (1.24) of a subnormal *bba* as a generalization of the Hartley entropy measure [39].

$$N(m) = \sum_{A \in 2^{\Omega} \setminus \emptyset} m(\mathbf{A}) \log_2 |A| + m(\emptyset) \log_2 |\Omega|, \tag{1.24}$$

where $0 \leq N(m) \leq \log_2 |\Omega|$.

A generalization of the nonspecificity in (1.24) can be deduced to obtained the global nonspecificity associated to the mass functions $m_1, \ldots, m_c$ with the following equation:

$$N(\{m_1, \ldots, m_c\}) = \frac{1}{n \log_2 |\Omega|} \sum_{k=1}^{c} N(m_k). \tag{1.25}$$

In evidential clustering (see Section 2.2.3 and Chapter 4), Equation (1.25) can be used to determine the optimal number of clusters by minimizing the global nonspecificity of the partition. In addition to Equation (1.24), several nonspecificity measures have been proposed in the literature such as Körner's [40] N^K and Yager's specificity N^Y defined respectively by

Equation (1.26) and (1.27). Recently, Yang et al. introduced in [41] a new nonspecificity measure based on belief intervals. The corresponding measure referred here to as N^{Ya} is defined in Equation (1.28).

$$N^K(m) = \sum_{A \subseteq \Omega} \log(m(A))|A|. \tag{1.26}$$

$$N^Y(m) = \sum_{A \subseteq \Omega / A \neq \emptyset} \frac{m(A)}{|A|}. \tag{1.27}$$

$$N^{Ya}(m) = \frac{1}{c} \sum_{k=1}^{c} (pl(\{\omega_k\}) - bel(\{w_k\})). \tag{1.28}$$

For detailed discussions on uncertainty measures in the framework of belief functions, we refer the readers to the recent review in [42].

Since Shafer's book publication "A mathematical theory of evidence" [9], the evidence theory has helped to overcome many problems. In the next section, we present some applications of the theory.

1.4.6 Applications of evidence theory

Among the applications of the evidence theory, artificial intelligence [43, 44], information fusion [45, 46], expert systems [47, 48], safety and reliability modeling [49], geographic information systems [50] can be cited. In clustering, the *evidential c-means* algorithm was introduced by Masson et al. in [51] as an extension of the *fuzzy c-means* algorithm. The extended method presented in Section 2.2.3, generates a partition called credal or evidential.

Despite its advantages, the evidence theory is also characterized by some limitations such as the use of large computational resources discussed in the following subsection.

1.4.7 Limitations of the evidence theory

One of the main drawbacks of the evidence theory holds on the need for large computational resources. Indeed, in clustering for instance the number of clusters when the evidence theory is used can grow exponentially due to the 2^c possible subsets of clusters (see the complexity analysis in Section 2.2.3 and 4.3). To limit the complexity of the methods, the size of the subsets can be limited to a given number. A detailed discussion of this solution is presented in Chapter 4.

Theories	Some advantages	Some limitations	Some properties
Fuzzy sets	Extension of hard sets theory, model imprecise data, fuzzy boundaries.	Interpretation of membership degrees, capturing overlapping clusters, assignment of crossover objects in clustering (see Remark 3).	Commutative, associative, transitive, De Morgan's laws.
Belief functions	Extension of hard and fuzzy sets theories, model imprecision and uncertainty of objects assignment, lower and upper bounds of *bba*, can capture overlapped clusters, conversion of *bba* to probabilities.	Large computational resources (see Section 2.2.3 and 4.3).	Monotone, equivalent representation of mass, belief, and plausibility functions.

Table 1.4: Summary of fuzzy sets, and belief functions advantages, limitations, and properties.

Summary

In this chapter, we present the probability, fuzzy sets, and evidence theories as uncertainty theories. We first discuss the hard sets theory as the basics of the three theories. For the probability theory, we focus our discussion on Shannon's entropy and presented some of its applications of this measure in clustering. In Section 1.2.1, 1.3 and 1.4 we respectively present the mathematical formulation and some properties of Shannon's entropy, fuzzy sets and belief functions.

In Table 1.4, we provide a summary of some advantages, limitations, and properties of fuzzy sets and belief functions in general and in clustering.

Key points

- The probability, fuzzy sets, and Dempster-Shafer theories are mathematical frameworks for modeling and reasoning with imprecision, vagueness, and uncertainty in data.
- Fuzzy sets are extensions of hard sets that allow gradual assignments of objects.
- Mass, belief, and plausibility functions are equivalent representations to quantify the degree of evidence of the state of a system.

2 — Clustering algorithms

Contents

2.1 Types of clustering partitions 46
2.1.1 Hard c-partitions 46
2.1.2 Fuzzy c-partitions 48
2.1.3 Evidential partitions 50
2.2 Partitioning algorithms 52
2.2.1 k-means 52
2.2.2 Fuzzy c-means 53
2.2.3 Evidential c-means 55
2.2.4 Entropy in fuzzy clustering 59
2.2.5 The need of categorical clustering algorithms 59
2.2.6 k-modes 62
2.2.7 Fuzzy k-modes 62
2.2.8 Fuzzy clustering with fuzzy centers 64

Introduction

This chapter aims to present some numerical and categorical partitioning clustering algorithms in the frameworks of hard sets, fuzzy sets, and Dempster-Shafer's theory of evidence. We first describe hard and fuzzy c-partitions and evidential partitions, then we provide examples of methods from which these partitions can be generated.

Definition 2.0.1 (Clustering)**.** In pattern recognition, cluster analysis or clustering is a family of unsupervised methods whose goal consists in finding hidden structures from high dimensional unlabeled data. The algorithms usually generate partitions of the data called clusters such that the objects in the same clusters are more similar than the objects in different clusters.

In the literature, three main clustering families can be distinguished: hierarchical, density-based, and partition-based. In hierarchical clustering data objects are clustered using a hierarchy of clusters by either agglomerative (bottom-up) or divisive (top-down) strategies. Density-based clustering is a spatial clustering method that groups data based on the densest regions. In partition-based clustering methods, objects are partitioned into different clusters. This strategy also called objective function clustering optimizes a generalized sum of squared errors.

Let $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_i}, \ldots, \mathbf{x_n}\}$ be a set of n observations where each object is described by p attributes $(F_1, \ldots, F_l, \ldots, F_p)$. Let c be the number of desired clusters, $\mathbf{V} = (\mathbf{v_1}, \ldots, \mathbf{v_k}, \ldots, \mathbf{v_c})$ the centers of the clusters, $\mu_{\mathbf{ik}}$ and $d(\mathbf{x_i}, \mathbf{v_k}) = d_{ik}$ be respectively the membership degree and distance between the i_{th} object and the k_{th} cluster center.

The generalized sum of squared errors referred to as J is given by the following equation

$$J = \sum_{i=1}^{n} \sum_{k=1}^{c} \mu_{ik}^{\beta} d_{ik}^2, \tag{2.1}$$

where β is a positive coefficient.

In the literature, the most frequent clustering family encountered is the partition-based family. The reason behind this can be explained by the extensive studies of these methods since the 60s and their flexibility to be suited and extended to different types of data. Therefore this chapter will focus on the review of the following partition-based clustering methods: *k-means* and *fuzzy c-means* and some of their extensions to categorical data: *k-modes*, *fuzzy k-modes* and *fuzzy centers clustering*. For general reviews of clustering techniques, we refer the readers to the following survey papers [52, 53].

The type of data partitions in clustering depends on the theory used to express certainty, uncertainty, and imprecision. For instance, when the hard and fuzzy sets theories are used the partitions are called respectively

hard and fuzzy partitions. When the evidence theory is used, the partition is called evidential or credal. Detailed characteristics of each partition are discussed in the next subsection.

2.1 Types of clustering partitions

In a hard partition, an object can belong to at most one cluster whereas, in a fuzzy partition, objects are known to belong in one cluster, however, due to the imprecision of the assignments, the fuzzy sets theory is used to model the imprecision which allows objects to have a membership degree in each cluster. As the Dempster-Shafer theory of evidence is a generalization of hard and fuzzy sets theory, evidential partitions extend the hard and fuzzy partitions. In these partitions, similarly to fuzzy partitions, the evidence theory is used to model the uncertainty of object assignments to clusters.

Before presenting the hard, fuzzy and evidential partitions let's define the notion of a partition.

Definition 2.1.1 (Partition)**.** Let $\mathbf{S}$ be a set, a (hard) c-partition of $\mathbf{S}$ corresponds to c nonempty subsets of $\mathbf{S}$ such that all the subsets are pairwise disjoint and cover $\mathbf{S}$. That is, if $\mathbf{S}$ is finite ($|\mathbf{S}| = n$) and $\mathbf{A_k}, \forall k \in \{1, \ldots, c\}$ are the subsets of $\mathbf{S}$ then:

$$\mathbf{A_k} \neq \emptyset, \forall k, \tag{2.2a}$$

$$\mathbf{A_k} \cap A_j = \emptyset, \forall k \neq j, \tag{2.2b}$$

$$\bigcup_{k=1}^{c} \mathbf{A_k} = \mathbf{S}. \tag{2.2c}$$

■

Definition 2.1.1 can be extended to fuzzy sets to define fuzzy c-partitions. In the following subsection, we provide examples of hard c-partitions and applications of the definition to clustering.

2.1.1 Hard c-partitions

To better understand hard c-partitions, let's consider the case where the number of partitions denoted by c is 2, the partition is then called hard 2-partitions.

Hard 2-partitions

If $\mathbf{A}$ is a hard set and $\neg\mathbf{A}$ its complement, the pair $(\mathbf{A}, \neg\mathbf{A})$ is a hard 2-partition. Indeed, all Equation (2.2) are satisfied for the pair $(\mathbf{A}, \neg\mathbf{A})$. Following the dual representation of hard sets operations in Equation (1.3), for the pair $(\mathbf{A}, \neg\mathbf{A})$, Equation (2.2) becomes

$$0 < \sum_{i=1}^{n} \mu_A(x_i) < n, \;\; 0 < \sum_{i=1}^{n} \mu_{\neg A}(x_i) < n, \tag{2.3a}$$

$$\mu_A(x_i), \;\; \mu_{\neg A}(x_i) \in \{0, 1\} \;\forall i, \tag{2.3b}$$

$$\mu_A(x_i) + \mu_{\neg A}(x_i) = 1 \;\forall 1 \leq i \leq n. \tag{2.3c}$$

Example 2.1.1 (Hard 2-partitions)**.** Let $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \mathbf{x_4}\}$, $\mathbf{A} = \{\mathbf{x_1}, \mathbf{x_2}\}$ and $\mathbf{B} = \{\mathbf{x_3}, \mathbf{x_4}\}$. The pair $(\mathbf{A}, \mathbf{B})$ is a hard 2-partitions. Indeed, we have $\mathbf{B} = \neg\mathbf{A}$ and $\forall x \in X$ all the Equation (2.3) are satisfied.

A matrix representation of hard 2-partitions can be obtained by setting the subsets as columns and the objects as rows. The values of the matrix correspond to the membership degrees of objects.

Example 2.1.2 (Matrix representation of a hard 2-partitions)**.** The matrix representation of the hard 2-partitions in Example 2.1.1 is given by

Objects	**A**	**B**
$\mathbf{x_1}$	1	0
$\mathbf{x_2}$	1	0
$\mathbf{x_3}$	0	1
$\mathbf{x_4}$	0	1

Table 2.1: Matrix representation of the hard 2-partitions in Example 2.1.1

Equation (2.3) can be generalized to hard c-partitions with $c \geq 2$.

Hard c-partitions

Let $\mu_{ik} = \mu_k(\mathbf{x_i})$ be the membership degree of the i_{th} object to the k_{th} cluster. When $c \geq 2$, a generalization of Equations (2.3) to c-partitions is

given by the following equations:

$$0 < \sum_{i=1}^{n} \mu_{ik} < n, \quad \forall\, 1 \leq k \leq c, \tag{2.4a}$$

$$\mu_{ik} \in \{0, 1\} \quad \forall i, \forall k, \tag{2.4b}$$

$$\sum_{k=1}^{c} \mu_{ik} = 1 \quad \forall\, 1 \leq i \leq n. \tag{2.4c}$$

As μ_{ik} is binary, Equation (2.4c) means that each x_i is exactly in one subset of the c-subsets. Equation (2.4a) can be interpreted as no subset is empty and no subset is X.

From the generalized c-partitions conditions (2.4c) and (2.4a), a formal definition of hard c-partitions can be derived as follows:

Definition 2.1.2 (Hard c-partitions)**.** Let $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_i}, \ldots, \mathbf{x_n}\}$ be a finite set, c an integer, such that $2 \leq c < n$, μ_{ik} the membership degree of x_i to the set A_k, $U = [\mu_{ik}]$ the matrix containing μ_{ik} and V_{nc} the set of real $n * c$ matrices. A partition is said to be a hard c-partitions [26] if

$$M_{hp} = \left\{U \in V_{nc} \middle|\, \mu_{ik} \in \{0, 1\}; \sum_{k=1}^{c} \mu_{ik} = 1\ \forall i; 0 < \sum_{i=1}^{n} \mu_{ik} < n\ \forall k\right\}. \tag{2.5}$$

where M_{hp} is the set of all hard partitions matrices. ■

As extensions of hard c-partitions, fuzzy c-partitions definitions can be derived from the hard sets's.

2.1.2 Fuzzy c-partitions

We first consider the case when $c = 2$.

Definition 2.1.3 (Fuzzy 2-partitions [26])**.** Let $\mathcal{A}$ be a fuzzy set and $\neg\mathcal{A}$ its complement. The pair $(\mathcal{A}, \neg\mathcal{A})$ is a fuzzy 2-partition $\mathbf{X}$ if

$$\mathcal{A} \neq \emptyset \text{ and } \neg\mathcal{A} \neq \emptyset, \tag{2.6a}$$

$$\mathbb{0} < \mu_{\mathcal{A}} < \mathbb{1}. \tag{2.6b}$$

■

Remark 2. 1) Equations (2.3) and (2.6) are equivalent. 2) the difference between the hard and fuzzy 2-partitions holds on the fact that $\mu_{\mathbf{A}}$ and $\mu_{\neg\mathbf{A}}$ $\in \{0, 1\}$ whereas $\mu_{\mathcal{A}}$ and $\mu_{\neg\mathcal{A}} \in [0, 1]$.

Example 2.1.3 (Fuzzy 2-partitions)**.** Let $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \mathbf{x_4}\}$, let $\mathcal{A} = \{0.3/\mathbf{x_1}, 0.6/\mathbf{x_2}, 0.5/\mathbf{x_3}\}$ and $\mathcal{B} = \{0.7/\mathbf{x_1}, 0.4/\mathbf{x_2}, 0.5/\mathbf{x_3}, 1/\mathbf{x_4}\}$ be two fuzzy sets. The pair $(\mathcal{A}, \mathcal{B})$ is a fuzzy 2-partition and the matrix representation is given by

Objects	$\mathcal{A}$	$\mathcal{B}$
$\mathbf{x_1}$	0.3	0.7
$\mathbf{x_2}$	0.6	0.4
$\mathbf{x_3}$	0.5	0.5
$\mathbf{x_4}$	0	1

Table 2.2: Matrix representation of the fuzzy 2-partition.

While $\mathbf{x_4}$ is a full member of $\mathcal{B}$, $\mathbf{x_1}$ and $\mathbf{x_2}$ are partially members of both $\mathcal{A}$ and $\mathcal{B}$. As the membership degree of $\mathbf{x_3}$ is 0.5, $\mathbf{x_3}$ is called a crossover point.

Remark 3. A hard partition can be deduced from a fuzzy partition by applying the maximum principle rule (i.e., each object is assigned to the subset with the maximum membership degree). For instance the Table 2.2 can be converted into a hard 2-partition as follow:

Objects	$\mathcal{A}$	$\mathcal{B}$
$\mathbf{x_1}$	0	1
$\mathbf{x_2}$	1	0
$\mathbf{x_3}$	1	0
$\mathbf{x_4}$	0	1

Table 2.3: Hard 2-partition generated from the fuzzy partition Table 2.2.

It can be noted that when converting a fuzzy partition containing crossover points (i.e., $x/\mu(x) = 0.5$) to a hard partition, these points can be miss-assigned to a subset due to the maximum principle. For instance, $\mathbf{x_3}$ can be assigned to both $\mathcal{A}$ and $\mathcal{B}$. Depending on the application, for example in medical science, the miss-assignment can have serious consequences.

Similarly to hard c-partitions, a fuzzy c-partition can be defined as follows:

Definition 2.1.4 (Fuzzy c-partitions [26])**.** A matrix U is said to be a fuzzy c-partition of X if

$$M_{fp} = \left\{U \in V_{nc} \middle| \mu_{ik} \in [0,1]; \sum_{k=1}^{c} \mu_{ik} = 1 \ \forall i; 0 < \sum_{i=1}^{n} \mu_{ik} < n \ \forall k\right\}. \tag{2.7}$$

where M_{fp} is the set of all fuzzy c-partitions matrices. ■

In the following section, we define evidential partitions.

2.1.3 Evidential partitions

Definition 2.1.5 (Evidential partition)**.** Let $\Omega = \{\omega_1, \ldots, \omega_k, \ldots, \omega_c\}$ be the frame of discernment (sets of clusters) and $\mathbf{A_k}$ be a subset of Ω.

Let $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_i}, \ldots, \mathbf{x_n}\}$ be a collection of n objects. The matrice $M = [m_{ij}]$ is called evidential or credal partition of Ω if for any object i, Equation (1.12) is satisfied, i.e, if V_{n2^c} is the set of real $n * 2^c$ matrices, the following equation should be satisfied:

$$M_{ep} = \left\{M \in V_{n2^c} \middle| m_{ik} \in [0,1]; \sum_{A_k \subseteq \Omega} m_{ik} = 1 \ \forall i\right\}. \tag{2.8}$$

where M_{ep} is the set of all evidential partitions matrices. ■

Example 2.1.4 (Evidential partition)**.** Let $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \mathbf{x_4}\}$ be a collection of 4 objects. For $c = 2$ we have $\Omega = \{\omega_1, \omega_2\}$ and $2^\Omega = \{\emptyset, \{\omega_1\}, \{\omega_2\}, \Omega\}$. The Table 2.4 is an evidential partition.

	$\emptyset$	$\{\omega_1\}$	$\{\omega_2\}$	Ω
$\mathbf{x_1}$	1	0	0	0
$\mathbf{x_2}$	0	0.2	0.8	0
$\mathbf{x_3}$	0	0	0	1
$\mathbf{x_4}$	0	1	0	0

Table 2.4: Example of an evidential partition.

From Section 1.4, the following interpretations of the membership degrees of an evidential partition can be deduced:

- An object is said to belong with certainty to a cluster when its membership degree is equal to 1. For example, the objects $\mathbf{x_1}$ and $\mathbf{x_4}$ from the Example 2.1.4 respectively belong with certainty to the subset $\emptyset$ and $\{\omega_1\}$.
- When an object belongs with certainty to the empty set, there is strong evidence that the class of the object does not lie in the frame of discernment. It is the case of the object $\mathbf{x_1}$ in the Example 2.4. In clustering, the empty set class usually represents outliers.
- When an object belongs with certainty to the frame of discernment, the class of the latter can not be precisely determined as it can belong to all the subsets. For instance, the class of object 3 in Example 2.4 corresponds to complete ignorance.

In the following subsection, we describe the link between hard, fuzzy, and evidential partitions.

Link between hard, fuzzy, and evidential partitions

As explained in Section 1.3, a fuzzy partition can be transformed to a hard partition by applying the maximum principle rule. Similarly, an evidential partition can be transformed into a fuzzy partition by applying transformation procedures of mass functions. The link between the three partitions is as follows:

- When the *bba*s is certain and restricted to singletons, the evidential partition corresponds to a hard partition and the class of each object is known with certainty.
- When the *bba* of the objects are Bayesian and restricted to singletons, the evidential partition corresponds to a fuzzy partition.
- When there is no equivalence between the evidential, fuzzy, and hard partitions, a normalization of the evidential partition can be performed as discussed in Section 1.4 then, the pignistic transformation $BetP$ in (1.23) can be used to convert an evidential partition to a fuzzy. The transformed partition can then be used to obtain a hard partition.

In addition to the preceding transformations of an evidential partition, the latter can also be transformed into other types of partitions such as the possibilistic [54] and rough [55] partitions. For detailed discussions on these transformations, we refer the readers to [56].

In the following section, we describe some clustering algorithms in the frameworks of hard sets, fuzzy sets, and evidence theory.

2.2 Partitioning algorithms

In the literature, several partitioning clustering algorithms have been proposed. Among the methods, the *k-means*, and *fuzzy c-means* algorithms described below are the most famous hard and fuzzy methods for clustering numerical data. Due to their limitations to numerical data, the two methods have been extended to other types of data such as categorical. The extensions are usually performed by adapting the distance (e.g. *k-modes* and *fuzzy k-modes* and centers of cluster representations (prototypes) (e.g. *fuzzy centers clustering*).

2.2.1 k-means

The *k-means* is a hard partitioning clustering algorithm of numerical data introduced independently in [28] and [29]. The term *means* refers to the means of objects in each cluster used as centers.

Objective function

The cost function of the *k-means* algorithm is defined by Equation (2.1) where $\beta = 1$:

$$J_{KM}(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^{n} \sum_{k=1}^{c} \mu_{ik} d_{ik}^2, \tag{2.9}$$

The latter cost function should be optimize under the constraints in Equation (2.5). In general the distance used in (2.9) is the Euclidean distance.

Optimization

The optimization problem can be solved using an alternate optimization scheme proposed in [28]. It consists of fixing the variable let say $\mathbf{U}$ and to minimize $\mathbf{V}$ and inversely.

Let $\mathbf{U} = (\mu_1, \ldots, \mu_k, \ldots, \mu_c)$ be the hard c-partitions obtained with the *k-means* and Λ_k be the number of objects in the k_{th} partition.

When V is fixed, objects x_i are assigned to their nearest cluster, i.e., their membership degrees are given by:

$$\mu_{ik} = \begin{cases} 1 & d_{ik} = d_r \text{ with } r = \underset{o \in \{1,\ldots,c\}}{\arg\min}\ d_{io} \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

When $\mathbf{U}$ is fixed, a solution of $J_{KM}(\mathbf{v_k})$ is given by

$$\mathbf{v_k} = \frac{1}{\Lambda_k} \sum_{\mathbf{x_i} \in \mu_k} \mathbf{x_i}. \tag{2.11}$$

Algorithm and complexity analysis

The *k-means* algorithm is described in **Algorithm 1** as follows: in the first step, the centers are randomly initialized. Then the distance between objects $\mathbf{x_i}$ and centers $\mathbf{v}_k$ are computed, and each object is assigned to the nearest cluster using the Equation (2.10). In the next step, the centers are updated with the new means of the new clusters with Equation (2.11). Finally, the second step is repeated until convergence is reached (i.e., there is no improvement of the centers from an iteration to another).

Let $D = [d_{ik}]_{1 \leq i \leq n, 1 \leq k \leq c}$ be the distance matrix and T be the number of iterations of an algorithm until convergence.

The time complexity of Algorithm 1 can be computed by taking into account the number of operations in each step of the algorithm. The Euclidean distance requires ncp operations. For the updating of the partition matrix U, $O(nc)$ operations are needed. Finally, to update the centers of clusters, ncp operations are necessary. With T iterations until convergence, the time complexity of the *k-means* is given by $O(T(npc + nc + ncp))$ which corresponds asymptotically to $O(ncpT)$.

For the memory complexity, five main variables are used in Algorithm 1: $\mathbf{X}$, D, $\mathbf{U}$, $\mathbf{V}$ and J_{KM}. These variables require respectively np, nc, nc, cp and 1 memory space. Hence, the memory complexity of the *k-means* is given by $O(np + nc + cp)$.

2.2.2 Fuzzy c-means

To overcome the limitations of hard sets in the framework of clustering, applications of the fuzzy sets theory have been proposed in this framework. In [57], Dunn proposed in the *Fuzzy ISODATA* algorithm as a fuzzy extension of the *k-means*. As an improvement of Dunn's algorithm, Bezdek proposed in [26] the *fuzzy c-means* (*FCM*) algorithm.

Objective function

The objective function of *FCM* is defined by the cost function in Equation (2.1):

$$J_{FCM}(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^{n} \sum_{k=1}^{c} \mu_{ik}^{\beta} d_{ik}^2, \tag{2.12}$$

Algorithm 1 *k-means* algorithm (KM)

Require: $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_i}, \ldots, \mathbf{x_n}\}$ the numerical data, $2 \leq c < n$ the number of clusters, and $\epsilon \geq 0$ a threshold.

Output: $\mathbf{U}$ hard c-partitions of $\mathbf{X}$, and $\mathbf{V}$ the centers of clusters.

Begin
Randomly initialize $\mathbf{V}_0$.
$\tau \leftarrow 0$
repeat
$\tau \leftarrow \tau + 1$
Compute the distance matrix D with the Euclidean distance.
Update $\mathbf{U}_\tau = [\mu_{\mathbf{ik}}]_\tau$ with Equation (2.10).
Update $\mathbf{V}_\tau$ with Equation (2.11).
until $\|J_{KM}^{\tau-1} - J_{KM}^{\tau}\| \leq \varepsilon$
End

under the constraints (2.8). $\beta > 1$ corresponds to the fuzziness coefficient (i.e, as β increases, the partition becomes fuzzier).

Optimization

The optimization problem can be solved using an alternate optimization scheme as in *k-means*. Therefore, when $\mathbf{U}$ is fixed, $J_{FCM}(\mathbf{v}_k)$ is minimized iff

$$\mathbf{v_k} = \frac{\sum_{i=1}^{n} \mu_{ik}^{\beta} \mathbf{x_i}}{\sum_{i=1}^{n} \mu_{ik}^{\beta}}, \quad \forall k \in \{1, \ldots, c\}. \tag{2.13}$$

By considering the centers of clusters fixed, the updates of the membership degrees are given by

$$\mu_{\mathbf{ik}} = \frac{1}{\sum_{h=1}^{c} \left(\frac{d_{ik}^2}{d_{ih}^2}\right)^{\frac{1}{\beta-1}}}, \quad \forall i \in \{1, \ldots, n\}, \forall k \in \{1, \ldots, c\}. \tag{2.14}$$

The proof of Equations (2.13) and (2.14) are available in [26].

Algorithm and complexity analysis

Similar to the *k-means* algorithm, in the first step of *FCM*, the centers $\mathbf{V}$ are randomly initialized. Then the distance between the objects and centers is computed to update the partition matrix $\mathbf{U}$ with the Equation (2.14). This update is used to determine the new centers with Equation (2.13). The updating of the partition matrix and centers are repeated until the convergence

is reached. The *FCM* algorithm is summarized in **Algorithm** 2.

The time and memory complexity of the *FCM* algorithm can be determined in the same way as for the *k-means* algorithm. For the distance matrix D, ncp operations are needed for the Euclidean distance. For the partition matrix $\mathbf{U}$, c^2n operations are needed and for the centers of clusters $\mathbf{V}$, ncp operations are necessary. Therefore, the time complexity of Algorithm 2 is $O(T(ncp+c^2n))$. The time complexity of the centers updating can be reduced to $O(cn)$ [58], consequently, the overall time complexity of the *FCM* algorithm is in $O(T(ncp+cn)) \rightarrow O(ncpT)$. The memory complexity of this algorithm is the same as *k-means*'s as the size of the variables are the same.

Algorithm 2 *Fuzzy c-means* algorithm (FCM)

Require: $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_i}, \ldots, \mathbf{x_n}\}$ the numerical data, $2 \leq c < n$ the number of clusters, $\beta > 1$ weighting exponent, and $\epsilon \geq 0$ a threshold.
Output: $\mathbf{U}$ fuzzy c-partitions of $\mathbf{X}$, and $\mathbf{V}$ the centers of clusters.

```
Begin
Randomly initialize V_0.
τ ← 0
repeat
    τ ← τ + 1
    Compute the distance matrix D with the Euclidean distance.
    Update U_τ = [μ_ik]_τ with Equation (2.14).
    Update V_τ with Equation (2.13).
until ||J_FCM^{τ-1} − J_FCM^τ|| ≤ ε
End
```

2.2.3 Evidential c-means

The evidential c-means (*ECM*) is a numerical data clustering method introduced in [59] is an extension of the *FCM* algorithm in the framework of the evidence theory. The method uses mass functions as membership degrees of objects and generates evidential partition.

Objective function

Since any subsets $\mathbf{A}_k$ of $\Omega = \{\omega_1, \ldots, \omega_k, \ldots, \omega_c\}$ can be a focal set, the *ECM* algorithm represents not only centers for clusters but also centers for subsets with cardinality greater than one. For the latter, the authors in [59] proposed to associate to each nonempty subset $\mathbf{A}_k$ with cardinality greater

than one, the barycenter of singletons defined by the following equation:

$$\mathbf{v_k} = \frac{1}{|\mathbf{A}_k|} \sum_{\nu=1}^{c} s_\nu \mathbf{v}_\nu, \tag{2.15}$$

where $s_\nu = \begin{cases} 1 & \text{if} \quad \omega_\nu \in \mathbf{A}_k, \\ 0 & \text{otherwise.} \end{cases}$ and v_ν $\forall 1 \leq \nu \leq c$ the centers of singletons.

The objective function inspired from the noise clustering [60] is given the following equation:

$$J_{ECM}(\mathbf{M}, \mathbf{V}) = \sum_{i=1}^{n} \sum_{\mathbf{A}_k \subseteq \Omega} |\mathbf{A}_k|^\alpha m_{ik}^\beta d_{ik}^2 + \sum_{i=1}^{n} \rho^2 m_{i\emptyset}^\beta \tag{2.16}$$

such that, for all $i = \{1, \ldots n\}$ and for all $\mathbf{A}_k \subseteq \Omega$,

$$\sum_{\mathbf{A}_k \subseteq \Omega, \mathbf{A}_k \neq \emptyset} m_{ik} + m_{i\emptyset} = 1, \qquad m_{ik} \geq 0. \tag{2.17}$$

where d_{ik} corresponds to the Euclidean distance between the i_{th} object and the k_{th} center. The mass $m_{i\emptyset}$ denotes the mass of $\mathbf{x}_i$ allocated to the empty set, $\rho > 0$ is a fixed parameter allowing the user to control the importance given to the empty set i.e. ρ corresponds to the distance to the noise cluster.

The weighting coefficient $|\mathbf{A}_k|^\alpha$, which corresponds to the cardinality of $\mathbf{A}_k$ as a power of α, allows the penalization of the allocation of belief to subsets with high cardinality. As in *FCM*, $\beta > 1$ corresponds to the fuzziness exponent: β close to 1 gives an evidential partition similar to a crisp partition, whereas β with a high value provides a partition where coefficients are equally distributed throughout the clusters. Usually, α is set to 1 and β to 2 for numerical data clustering.

Optimization

The cost function (2.16) under the constraint (2.17) can be solved by an alternate optimization scheme. When $\mathbf{V}$ is fixed, the updating formula of the evidential partition $\mathbf{M}$ can be decomposed into two parts: 1) the updating of the masses of nonempty subsets and 2) the updating of the masses corresponding to the empty set. The two updating formulas are respectively given by Equation (2.18) and (2.19). The proof can be found in [59].

$$m_{ik} = \frac{|\mathbf{A}_k|^{-\alpha/(\beta-1)} d_{ik}^{-2/(\beta-1)}}{\sum_{A_\nu \neq \emptyset} |A_\nu|^{-\alpha/(\beta-1)} d_{i\nu}^{-2/(\beta-1)} + \rho^{-2/(\beta-1)}}. \tag{2.18}$$

For $\mathbf{A}_k = \emptyset$, the mass is defined as:

$$m_{i\emptyset} = 1 - \sum_{A_k \neq \emptyset} m_{ik} \quad \forall i = 1, \ldots n. \tag{2.19}$$

When $\mathbf{M}$ is fixed, the updating formula of the centers of singletons correspond to the solution of the linear system in (2.20) [59]

$$I\mathbf{V} = \Pi, \tag{2.20}$$

where I and Π are respectively the matrices $(c \times c)$ (2.21) and (2.22) $(c \times p)$.

$$I_{\xi k} = \sum_{i=1}^{n} \sum_{A_k \supseteq \{\omega_\xi, \omega_k\}} |A_k|^{\alpha-2} m_{ik}^{\beta}, \quad \forall \xi, k = 1, c. \tag{2.21}$$

$$\Pi_{l\nu} = \sum_{i=1}^{n} \mathbf{x}_{il} \sum_{\omega_\nu \in A_k} |A_k|^{\alpha-2} m_{ik}^{\beta}, \quad \forall \nu = 1, c, \forall l = 1, p. \tag{2.22}$$

Algorithm and complexity analysis

The first step of *ECM*'s algorithm summarized in **Algorithm** 3 as for the *k-means* and *FCM* algorithms consists in randomly initializing the centers. Then the distances between objects and centers are used to update the evidential matrix $\mathbf{M}$ and the centers $\mathbf{V}$ respectively with Equation (2.18) for non-empty sets and (2.19) for the empty set and Equation (2.20). The updates are repeated until convergence.

As the *ECM* algorithm uses subsets of clusters, the time complexity is an exponential function of c, i.e., 2^c. To reduce this complexity, the number of subsets of clusters can be limited to the empty set, single clusters, and Ω, i.e., $c + 2$ subsets. Below we analyze the time and memory complexity of *ECM*.

Let ϱ be the number of desired subsets of clusters from *ECM*, for instance, $\varrho = c + 2$ or $\varrho = 2^c$ and let $O(Q)$ be the complexity of the linear system used to solve (2.20).

The time complexity of the *ECM* algorithm is given by $O(p^2 n\varrho T + QT)$. Indeed, the computation of the distance matrix needs $n\varrho p$ operations. For m_{ik}, for each i, the denominator in (2.18) requires ϱ operations to compute the sum. For n objects and ϱ subsets of clusters we have a time complexity $O(n\varrho)$ to update the evidential partition. For the matrices H and Π, $c^2 n\varrho$ and $cpn\varrho$ operations are respectively needed. Then the time complexity to update of the centers is $O(n\varrho p + n\varrho + c^2 n\varrho + cpn\varrho + Q) \rightarrow O(p^2 n\varrho + Q)$ by considering $p >> c$. Consequently, the overall time complexity of the *ECM*

algorithm is $O(p^2 n\varrho T + QT)$ with T the number of iterations until convergence.

The memory complexity of *ECM* is given by $O(np + n\varrho + p\varrho)$ as $\mathbf{X}$, D, $\mathbf{M}$ and $\mathbf{V}$ respectively needs np, $n\varrho$, $n\varrho$ and ϱp memory allocations.

Algorithm 3 *Evidential c-means* algorithm (ECM)

Require: $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_i}, \ldots, \mathbf{x_n}\}$ the numerical data, $2 \leq c < n$ the number of clusters, $\alpha \geq 1$ the weighting exponent for cardinality, $\beta > 1$ weighting exponent, and $\rho > 0$ the distance to the outliers cluster and $\epsilon \geq 0$ a threshold.

Output: $\mathbf{M}$ evidential partition of, $\mathbf{X}$, and $\mathbf{V}$ the centers of clusters.

Begin
Randomly initialize $\mathbf{M}_0$.
$\tau \leftarrow 0$
repeat
$\tau \leftarrow \tau + 1$
Compute the distance matrix D with the Euclidean distance.
Update $\mathbf{V}_\tau$ with the solution of Equation (2.20).
Update $\mathbf{M}_\tau = [m_{ik}]_\tau$ with Equation (2.18) and (2.19).
until $\|J_{ECM}^{\tau-1} - J_{ECM}^{\tau}\| \leq \varepsilon$
End

Shannon's entropy described in subsection 2.2.4 has been applied to partitioning clustering algorithms particularly in fuzzy clustering. The following subsection presents some of these applications.

2.2.4 Entropy in fuzzy clustering

Entropy-based fuzzy clustering methods are extensions of fuzzy clustering in which a weighted entropy is incorporated into the objective function. Depending on the application, the entropy can have different roles and meanings. In [61, 62], the authors proposed a fuzzy clustering method in which the entropy is seen as a regularizing function to the objective function of *FCM*. While in [61] the authors use the entropy to propose a new generalization of the *k-means*, in [62], the entropy is used to maximize the dissimilarity between clusters. In [63], the authors used the entropy as a prior in Bayesian context for image restoration and proposed a new clustering method based on the fuzzy framework. In [64], an entropy-based fuzzy clustering method that automatically identifies the number and initial locations of cluster centers is proposed. In [65], the entropy of the membership functions is incorporated into the objective function to allow a gradual transition from a maximum uncertainty to a minimum uncertainty during the clustering process. When applied for clustering validation, the entropy corresponds to an internal validity index [26]. In this case, the entropy measures the fuzziness of partitions produced by clusters.

The *k-means*, *FCM* and *ECM* algorithms described previously only handle numerical data. However, real-world data is not limited to numerical. Non-numerical data transformations are usually used in order to employ numerical clustering methods. For instance, categorical data can be transformed into numerical using encoding techniques. However, these transformations have drawbacks. We present and discuss in the following subset some limitations of two popular encoding techniques namely *label encoding* and *one-hot encoding*.

2.2.5 The need of categorical clustering algorithms

Let $\mathrm{Dom}(F_l) = \{a_l^{(1)}, \ldots, a_l^{(t)}, \ldots, a_l^{(n_l)}\}$ be the domain of attributes F_l $\forall l$ of $\mathbf{X}$. The label and one-hot encoding techniques are performed as follows.

Label encoding

The *label encoding* method consists in assigning a numerical code to each category $a_l^{(t)}$ of F_l $\forall t$ and $\forall l$.

Example 2.2.1 (Label encoding)**.** Let F be a categorical variable and $Dom(F)$ = { Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday }. Using the *label encoding* technique, the categories Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday can be respectively encoded $1, 2, 3, 4, 5, 6, 7$.

Limitations of label encoding

Despite the simplicity of the *label encoding* method for converting categorical data to numerical, it has some limitations. For instance, operations can be applied on the transformed data that do not make sense when transposed to the original data. For example, operations $1 + 2$, $7 - 3$, $6/2$ or $3 * 5$ can be performed, but Monday $-$ Tuesday or Wednesday $*$ Friday have no meaning. Hence the distances used for instance in *k-means FCM* and *ECM* algorithms that are based on operations of the variables is inappropriate on the transformed data with *label encoding*.

One-hot encoding

Similarly to the *label encoding*, the *one-hot encoding* is a technique for transforming categorical data to numerical. The method encodes each category to binary with 1 if the category appear and 0 otherwise.

Example 2.2.2 (One-hot encoding)**.** Let A_1 and A_2 be two categorical variables and $Dom(A_1) = \{yes, no\}$, $Dom(A_2) = \{high, medium, small\}$ be their respective domains. Let consider the following table

Objects	A_1	A_2
$\mathbf{x_1}$	yes	medium
$\mathbf{x_2}$	no	high
$\mathbf{x_3}$	yes	small

Table 2.5: Categorical data to *one-hot encode*.

The corresponding *one-hot encoded* data is given by

Objects	A_1(yes)	A_1(no)	A_2(high)	A_2(medium)	A_2(small)
$\mathbf{x_1}$	1	0	0	1	0
$\mathbf{x_2}$	0	1	1	0	0
$\mathbf{x_3}$	1	0	0	0	1

Table 2.6: *One-hot encoding* of Table 2.5.

Limitations of one-hot encoding

It can be noted that in Table 2.6 as n_l increases the dimension of the *one-hot encoded* data increases too. For instance, if $\mathbf{X}$ is a categorical data containing p attributes and each attribute of $\mathbf{X}$ is described by n_l categories then the dimension of $\mathbf{X}^{one\text{-}hot}$ ($\mathbf{X}$ *one-hot encoded*) is $\sum_{l=1}^{p} n_l$. Therefore, the *one-hot encoding* technique is suited only for categorical data containing small number of categories.

The limitations of the *label encoding* and *one-hot encoding* show that clustering algorithms that can handle categorical data without transformation are needed especially for data with high categories per attribute. To adapt numerical clustering methods to categorical data, categorical data dissimilarity measures can be used. In the literature, several categorical measures have been proposed (see [66] for a review). Among the measures, the most used is the Hamming distance [67] also called simple matching dissimilarity measure. It is defined as follows:

$$d^H(\mathbf{x_i}, \mathbf{x_k}) = \sum_{l=1}^{p} \delta(\mathbf{x_{il}}, \mathbf{x_{kl}}), \tag{2.23}$$

where

$$\delta(\mathbf{x_{il}}, \mathbf{x_{kl}}) = \begin{cases} 0 & \text{if} \quad \mathbf{x_{il}} = \mathbf{x_{kl}}, \\ 1 & \mathbf{x_{il}} \neq \mathbf{x_{kl}}. \end{cases} \tag{2.24}$$

Hence, the Hamming distance corresponds to the number of mismatching categories of two objects $\mathbf{x_i}$ and $\mathbf{x_k}$.

The Hamming distance has been used in several categorical clustering methods such as the *k-modes* [27], the *fuzzy k-modes* [68] and in *fuzzy centers clustering* [30] of categorical data. The following sections describe each method.

2.2.6 k-modes

The *k-modes* [27] algorithm is an extension of *k-means* for clustering categorical data. The method uses the Hamming distance as a dissimilarity measure between objects and cluster centers. The term *modes* refer to the use of modes of objects as centers instead of means in the *k-means* algorithm.

Definition 2.2.1 (Mode)**.** A mode of $\mathbf{X}$ described by p categorical attributes is a vector $Q = [q_1, \ldots, q_l, \ldots, q_p]$ that minimizes

$$d(X, Q) = \sum_{i=1}^{n} d^H(\mathbf{x_i}, Q). \tag{2.25}$$

Objective function

The objective function of the *k-modes* remains the same as the *k-means*'s by setting the squared Euclidean distance by Hamming's and the centers to the modes of attributes in each cluster.

Optimization

The objective function of the *k-modes* can be optimized similarly to the *k-means*. To constitute the hard partition matrix $\mathbf{U}$, at each iteration, objects are assigned to the closest centers by on the Hamming distance. The update of the centers is performed by determining the modes of attributes in each cluster with Equation (2.25).

Algorithm and complexity analysis

The *k-modes* algorithm can be derived from Algorithm 1. It is described in **Algorithm** 6. For the time complexity, the algorithm requires ncp operations to compute the Hamming distance. The computation of the centers of clusters, i.e., the modes, requires ncp operations, and the update of the partition matrix nc operations. Consequently, the overall time complexity of this algorithm is the same as *k-means*'s with the Euclidean distance. Similarly, the memory complexity remains the same as the size of variables as the same.

2.2.7 Fuzzy k-modes

As discussed in Section 2.2.1, the *k-means* algorithm and the *k-modes* are limited when expressing fuzziness. To overcome this limitation, Huang proposed an extension of the *k-modes* algorithm to the fuzzy framework. The

Algorithm 4 *k-modes* algorithm

Require: $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_i}, \ldots, \mathbf{x_n}\}$ the categorical data, $2 \leq c < n$ the number of clusters, and $\epsilon \geq 0$ a threshold.

Output: $\mathbf{U}$ hard c-partitions of $\mathbf{X}$ and $\mathbf{V}$ the clusters centers.

Begin

Randomly initialize $\mathbf{V}_0$.

$\tau \leftarrow 0$

repeat

$\tau \leftarrow \tau + 1$

Compute the distance matrix D with the Hamming distance.

Update $\mathbf{U}_\tau = [\mu_{\mathbf{ik}}]_\tau$ with Equation (2.10) where $d_{ik} = d_{ik}^H$.

Update $\mathbf{V}_\tau$ by determining the modes of attributes in each cluster with Equation (2.25).

until $\|J_{k-modes}^{\tau-1} - J_{k-modes}^{\tau}\| \leq \varepsilon$

End

new algorithm referred to as *fuzzy k-modes* (*FKM*)[68] is an adaptation of the *FCM* algorithm. Similar to the adaptation of the *k-means* algorithm to the *k-modes*, the *FKM* algorithm uses modes of objects instead of means to represent the centers of the clusters and the Hamming distance as a dissimilarity measure.

Objective function

The objective function of the *FKM* algorithm remains the same as *FCM* with $d_{ik}^2 = d_{ik}^H$.

Optimization

The updating formula can be obtained by using an alternate optimization scheme as for the *FCM*. Therefore when the centers $\mathbf{V}$ are fixed, the updating formula of the membership degrees μ_{ik} is given by Equation (2.14). When the partition matrix U is fixed, the updating of the centers $\mathbf{v_k}$ are given by: $\mathbf{v_{kl}} = a_l^{(r)} \in \mathrm{DOM}(F_l)$ where

$$r = \underset{t \in \{1,\ldots,n_l\}}{\arg\max} \sum_{i/\mathbf{x_{il}} = \mathbf{a}_\mathbf{l}^{(\mathbf{t})}} \mu_{ik}^{\beta} \tag{2.26}$$

Equation (2.26) denotes that the centers correspond to the categories of F_l $1 \leq l \leq p$ having the highest frequency $\sum_{i,\mathbf{x_{ik}}=a_l^{(t)}} \mu_{ik}^{\beta}$, in other words the centers correspond to the *modes* of objects.

Algorithm and complexity analysis

The *FKM* algorithm can be derived from *FCM* algorithm is as follows

Algorithm 5 *Fuzzy k-modes* algorithm (FKM)

Require: $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_i}, \ldots, \mathbf{x_n}\}$ the categorical data, $2 \leq c < n$ the number of clusters, and $\epsilon \geq 0$ a threshold.

Output: $\mathbf{U}$ fuzzy c-partitions of $\mathbf{X}$ and $\mathbf{V}$ the clusters centers.

Begin
Randomly initialize $\mathbf{V}_0$.
$\tau \leftarrow 0$
repeat
$\tau \leftarrow \tau + 1$
Compute the distance matrix D with the Hamming distance.
Update $\mathbf{U}_\tau = [\mu_{\mathbf{ik}}]_\tau$ with Equation (2.14) where $d^2_{ik} = d^H_{ik}$.
Update $\mathbf{V}_\tau$ Equation (2.26).
until $\|J^{\tau-1}_{FKM} - J^{\tau}_{FKM}\| \leq \varepsilon$
End

The time and memory complxity of the *FKM* algorithm remains the same as *FCM*'s, i.e., they are respectively $O(ncpT)$ and $O(np + nc + cp)$.

Remark 4 (A limitation of frequency-based clustering methods)**.** It can be noticed that the frequency-based methods (here *k-modes* and the *FKM*) can be limited when the frequency of the attributes categories are close to each other or the same. For instance, let's consider the following example:

Example 2.2.3. Let F be a categorical attribute with DOM(F) = {*yes*, *no*}. Let's consider the frequency of the categories to be 50 and 49 respectively for *yes* and *no*. The *k-modes* based methods will consider the category *yes* as representative of the attribute F even though the category *no* has a similar frequency. When the frequencies are the same, a category will be arbitrarily chosen to be the representative of the attribute which can harm the performance of the method, and depending on the application the miss-representation of the clusters can be problematic.

To overcome this limitation a new representation of cluster centers was proposed in [30]. The authors introduced a fuzzy representation of the cluster centers by using fuzzy sets.

2.2.8 Fuzzy clustering with fuzzy centers

The *k-modes* and *FKM* clustering methods are hard centers based clustering methods, i.e., as they use modes of attributes, only one category per attribute is considered. In contrast, fuzzy centers (FC) introduced in

[30] use the fuzzy sets theory by associating a weight to each attribute category representing its membership degree to the corresponding centers (fuzzy sets). With this representation, all the attribute categories can contribute to the centers and their weights indicate their importance. More formally, if $\mathbf{v_k} = (\mathbf{v_{k1}}, \ldots, \mathbf{v_{kl}}, \ldots, \mathbf{v_{kp}})$ is the center of the k_{th} cluster and $\mathbf{v_{kl}}$ the fuzzy set $(a_l^{(t)},\ w_{kl}^{(t)})\ \forall l \in \{1, \ldots, p\},\ \forall t \in \{1, \ldots, n_l\}$ corresponding to the center associated to the l_{th} attribute, then $\mathbf{v_{kl}}$ is defined by:

$$\mathbf{v_{kl}} = \{w_{kl}^{(1)}/a_l^{(1)}, \ldots, w_{kl}^{(t)}/a_l^{(t)}, \ldots, w_{kl}^{(n_l)}a_l^{(n_l)}\} \tag{2.27}$$

with

$$0 \leq w_{kl}^{(t)} \leq 1, \quad 1 \leq t \leq n_l \tag{2.28}$$

$$\sum_{t=1}^{n_l} w_{kl}^{(t)} = 1, \quad 1 \leq l \leq p. \tag{2.29}$$

Equation (2.27) therefore corresponds to a generalization of hard centers such as in the *k-modes* and *FKM* algorithms and overcome their limitation when the frequencies of the categories are close to each other or the same. In this equation, the weights $w_{kl}^{(t)}$ are positive, less than 1 (2.28) and sum up to 1 (2.29).

Example 2.2.4 (Fuzzy centers)**.** Let A_1 and A_2 be two categorical attributes with $DOM(A_1) = DOM(A_2) = \{a_1^{(1)},\ a_1^{(2)}\} = \{yes,\ no\}$. Let's consider the following data set:

	A_1	A_2
$\mathbf{x_1}$	yes	no
$\mathbf{x_2}$	no	yes
$\mathbf{x_3}$	yes	no

Table 2.7: Categorical data set to illustrate fuzzy centers.

- A fuzzy centroid $\mathbf{v_k}$ restricted to A_1 can be

$$\mathbf{v_{k1}} = \{0.6/yes, 0.4/no\}.$$

- A fuzzy center $\mathbf{v_k}$ restricted to A_2 can be

$$\mathbf{v_{k2}} = \{0.8/yes, 0.2/no\}.$$

- A fuzzy centers $\mathbf{v_k}$ of a cluster can be

$$\begin{aligned}\mathbf{v_k} &= (\mathbf{v_{k1}},\ \mathbf{v_{k2}}) \\ &= (\{0.6/yes, 0.4/no\},\ \{0.8/yes, 0.2/no\}).\end{aligned}$$

Objective function

The objective function of *FC* remains the same as *FCM*. Only the squared distance in the Equation (2.12) changes:

$$\begin{cases} J_{FC}(\mathbf{U},\mathbf{V}) = \sum_{i=1}^{n}\sum_{k=1}^{c}\mu_{ik}^{\beta} d_{ik}^{GH} \\ s.t.\ \mathbf{U} \text{ satisfies Equation (2.8) and } \mathbf{V} \text{ satisfies Equations (2.28) and (2.29).} \end{cases} \tag{2.30}$$

where d^{GH} is a generalized Hamming distance that takes into account category weights. It is defined as follows:

$$d_{ik}^{GH} = \sum_{l=1}^{p}\sum_{t=1}^{n_l} \delta(\mathbf{x_{il}}, a_l^{(t)}), \tag{2.31}$$

where

$$\delta(\mathbf{x_{il}}, a_l^{(t)}) = \begin{cases} 0 & \text{if} \quad \mathbf{x_{il}} = a_l^{(t)}, \\ w_{kl}^{(t)} & \text{if} \quad \mathbf{x_{il}} \neq a_l^{(t)}. \end{cases} \tag{2.32}$$

The new distance (2.31) is then the sum of weights of dissimilar categories between objects $\mathbf{x_i}$ and the centers $\mathbf{v_k}$. The following example illustrates this distance.

Example 2.2.5 (Generalized Hamming distance example)**.** Let's consider again the data set in Table 2.7 and the center $\mathbf{v_k}$ defined in the Example 2.2.4. Let's d_{1k}^{GH} be the distance between $\mathbf{v_k}$ and $\mathbf{x_1}$. Let $\delta_{1k}^{(1)}$ and $\delta_{1k}^{(2)}$ be respectively the relative distance between $\mathbf{x_{11}}$ and $\mathbf{v_{k1}}$ and $\mathbf{x_{12}}$ and $\mathbf{v_{k2}}$. We have

- We have $\mathbf{v_{k1}} = \{0.6/yes, 0.4/no\}$, and $\mathbf{x_{11}} = yes$. As $a_1^{(1)} = yes$ and $a_1^{(2)} = no$, from Equation (2.31) we add the weight of $a_1^{(2)}$ (i.e., $w_{k1}^{(2)}$) to $\delta_{k1}^{(1)}$ hence

$$\delta_{k1}^{(1)} = w_{k1}^{(2)}$$
$$\delta_{k1}^{(1)} = 0.4.$$

- We have $\mathbf{v_{k2}} = \{0.8/yes, 0.2/no\}$, and $\mathbf{x_{12}} = no$. As $a_2^{(1)} = yes$ and $a_2^{(2)} = no$, from Equation (2.31) we add the weight of $a_2^{(1)}$ (i.e., $w_{k2}^{(1)}$) to $\delta_{k1}^{(2)}$ hence

$$\delta_{k2}^{(2)} = w_{k2}^{(1)}$$
$$\delta_{k2}^{(2)} = 0.8.$$

Therefore we have

$$\begin{aligned} d_{1k}^{GH} &= \delta_{k1}^{(1)} + \delta_{k1}^{(2)} \\ d_{1k}^{GH} &= 0.4 + 0.8 \\ d_{1k}^{GH} &= 1.2. \end{aligned}$$

Optimization

When $\mathbf{V}$ is fixed, i.e., the weights $w_{kl}^{(t)}$ are fixed, the updating formula of the membership degrees μ_{ik} is the same as *FCM*'s with $d_{ik}^2 = d_{ik}^{GH}$. When $\mathbf{U}$ is fixed, the centers of clusters are updated by updating the weights associated to each attribute category. In [30], these updates are given by the following equation

$$w_{kl}^{(t)} = \sum_{i=1}^{n} \gamma(x_{il}), \tag{2.33}$$

where

$$\gamma(\mathbf{x_{il}}) = \begin{cases} \mu_{ik}^{\beta} & \text{if} \quad \mathbf{x_{il}} = a_l^{(t)}, \\ 0 & \text{if} \quad \mathbf{x_{il}} \neq a_l^{(t)}. \end{cases} \tag{2.34}$$

Algorithm and complexity analysis

As for the *FKM* algorithm, the FC algorithm can be derived from *FCM* algorithm is as follows

Algorithm 6 *Fuzzy center* algorithm (FC)

Require: $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_i}, \ldots, \mathbf{x_n}\}$ the categorical data, $2 \leq c < n$ the number of clusters, and ϵ a threshold.

Output: $\mathbf{U}$ fuzzy c-partitions of $\mathbf{X}$ and $\mathbf{V}$ the clusters centers.

Begin
Randomly initialize $\mathbf{V}_0$.
$\tau \leftarrow 0$
repeat
$\tau \leftarrow \tau + 1$
Compute the distance matrix D with the generalized Hamming distance.
Update $\mathbf{U}_\tau = [\mu_{\mathbf{ik}}]_\tau$ with Equation (2.14) where $d_{ik}^2 = d_{ik}^{GH}$.
Update $\mathbf{V}_\tau$ Equation (2.33).
until $\|J_{FC}^{\tau-1} - J_{FC}^{\tau}\| \leq \varepsilon$
End

Algorithms	Time complexity	Memory complexity
KM	$O(npcT)$	$O(np+nc+cp)$
FCM	$O(ncpT)$	$O(np+nc+cp)$
ECM	$O(p^2n\varrho T+QT)$	$O(np+n\varrho+\varrho p)$
k-modes	$O(npcT)$	$O(np+nc+cp)$
FKM	$O(ncpT)$	$O(np+nc+cp)$
FC	$O(nc\mathcal{J}T)$	$O(np+nc+c\mathcal{J})$

Table 2.8: Time and memory complexity of *KM*, *FCM*, *ECM*, *k-modes*, *FKM* and *FC* algorithms.

As the previous algorithms, the time complexity of the *FC* algorithm can be determine by determining the complexity of the computation of $\boldsymbol{D}$, $\mathbf{U}$, and $\mathbf{V}$. For the distance, $nc\mathcal{J}$ operations are needed where $\mathcal{J}=\sum_{l=1}^{p}n_l$. For $\mathbf{U}$ and $\mathbf{V}$ the time complexity are respectively $O(c^2n)$ and $O(cn\mathcal{J})$. By applying the time complexity reducing in [58], the complexity of $\mathbf{U}$ becomes $O(cn)$. The overall time complexity is then given by $O(nc\mathcal{J}T))$.

For the memory complexity, only the memory size of the centers of clusters change from the $\boldsymbol{FKM}$ algorithm. The new memory size needed to store $\mathbf{V}$ is $c\mathcal{J}$. Therefore, the overall memory complexity of the $\boldsymbol{FC}$ algorithm is $O(np+nc+c\mathcal{J})$.

In Chapter 3, we show that (2.33) may present convergence issues.

Summary

In this chapter, we present applications of the hard sets, fuzzy sets, and evidence theory in clustering. We define hard, and fuzzy c-partitions and the evidential partitions that are based on the three theories. We then present numerical clustering algorithms *k-means*, *FCM* and *ECM* for generating such partitions. We illustrate some limitations of two popular encoding techniques for transforming numerical data to categorical and introduced the Hamming distance which is used in the *k-modes* and *FKM* algorithms. For each of the described methods, the objective function, updating formula, algorithm and complexity analysis are given.

Summary of time and memory complexity of *KM*, *FCM*, *ECM*, *k-modes*, *FKM* and *FC* algorithms are reported in Table 2.8.

Key points

- Clustering partitions depend on the theory used. When the hard sets, fuzzy sets, and the evidence theories are used the partitions are linked and are respectively called hard, fuzzy and evidential.
- Conversions of numerical data to categorical can have drawbacks such as the increase of the data dimensions.
- The fuzzy sets theory can be used for cluster center representations in which each attribute category contributes to the centers.
- The time and memory complexity of the *ECM* algorithm can be an exponential function of the number of clusters.

Part II

Contributions

3 — Categorical fuzzy entropy c-means

Contents

3.1 Issues in *FC* . . . 72
3.2 New updates of weights in *FC* algorithm . . . 72
3.2.1 Entropy as regularization function . . . 76
3.3 *CFE*: fuzzy entropy c-means . . . 76
3.4 Experiments . . . 79
3.4.1 Datasets . . . 80
3.4.2 Evaluation criteria . . . 84
3.4.3 Experimental protocol . . . 87
3.4.4 Parameter settings . . . 87
3.4.5 Materials . . . 88
3.4.6 Results . . . 88
3.5 Strengths of *CFE* . . . 92
3.6 Limitations of *CFE* . . . 93
3.6.1 Optimal values of β and Ψ . . . 93
3.6.2 Category frequencies vs weights from *CFE* . . . 94
3.6.3 Interpretation of fuzzy membership degrees . . . 96

Introduction

Among the variants of the *FKM* algorithm, we presented in Chapter 2 the *FC* algorithm that uses the fuzzy sets theory for object assignments in clusters and the centers of the cluster representations. We derived the

objective function of this method and noted that the updating formula of the centers may present convergence issues. In this chapter, an extended version of the *FC* algorithm called *categorical fuzzy entropy c-means* and referred to as *CFE*). The new method uses Shannon's entropy to regularized the attribute category weights. Through the experiences on nine datasets having different characteristics (number of objects, classes, attributes, categories), we compare the performance of the new method to existing numerical and categorical clustering methods. Finally, we demonstrate the strengths of *CFE* and discuss some of its limitations.

3.1 Issues in *FC*

In [30], the authors introduced the fuzzy representation of categorical clusters centers using the fuzzy sets theory. It can be noticed that the updating of centers in Equation (2.33) presents some issues. We can note that the Equation (2.33) does not satisfy the constraints (2.28) and (2.29). Indeed, we have from Equation (2.33):

$$\begin{aligned} w_{kl}^{(t)} &= \sum_{i=1}^{n} \gamma(\mathbf{x_{il}}) \\ &= \sum_{i,\mathbf{x_{il}}=a_l^{(t)}} \mu_{ik}^{\beta}, \end{aligned}$$

hence we have

$$\sum_{t=1}^{n_l} w_{kl}^{(t)} = \sum_{t=1}^{n_l} \sum_{i,\mathbf{x_{il}}=a_l^{(t)}} \mu_{ik}^{\beta} \neq 1. \tag{3.1}$$

Equation (3.1) shows that neither the constraint (2.28) nor (2.29) are satisfied. Therefore the *FC* algorithm may not converge. To overcome this issue, we derived the objective function *FC*. The new updating formula of the centers is presented in the next section.

3.2 New updates of weights in *FC* algorithm

To rigorously determine the updating formula of the weights in *FC* algorithm, we derived the objective and came up with the following theorem:

Theorem 3.2.1. Let $S_{kl}^{(t)}$ be the frequency corresponding to the t_{th} category of the l_{th} attribute in the k_{th} cluster, i.e., $S_{kl}^{(t)}$ is defined by

$$S_{kl}^{(t)} = \sum_{i,x_{il}=a_l^{(t)}} \mu_{ik}^{\beta}.$$

Let $\tau_1, \ldots, \tau_{q-1}$ be the indices of attribute categories where the frequency of the t_{th} category equals the frequencies of the categories $\tau_1, \ldots, \tau_{q-1}$ such that $t \neq \tau_1 \neq \ldots \neq \tau_{q-1}$, i.e., $S_{kl}^{(t)} = S_{kl}^{(\tau_1)} = \ldots = S_{kl}^{(\tau_{q-1})}$, then when $\mathbf{U}$ fixed the objective function (2.30) is minimized iff

$$w_{kl}^{(t)} = \begin{cases} 1 & \text{if } S_{kl}^{(t)} = S_{kl}^{(r)} \text{ with } r = \underset{\nu \in \{1,\ldots,n_l\}}{\arg\max}\, S_{kl}^{(\nu)} \\ \dfrac{1}{q} & \text{if } \exists \tau_1, \ldots, \tau_{q-1} \; s.t. \; S_{kl}^{(r)} = S_{kl}^{(\tau_1)} = \ldots = S_{kl}^{(\tau_{q-1})} > S_{kl}^{(t)}, \\ & \forall r, t \in \{1, \ldots, n_l\} \text{ with} \\ & r \neq \tau_1 \neq \ldots \neq \tau_{q-1} \neq t. \\ 0 & \text{otherwise} \end{cases}. \tag{3.2}$$

Proof. The objective function J_{FC} with the squared distance replaced by (2.31) can be rewritten as

$$\begin{aligned} J_{FC}(\mathbf{U}, \mathbf{V}) &= \sum_{i=1}^{n} \sum_{k=1}^{c} \mu_{ik}^{\beta} d_{ik}^{GH} \\ &= \sum_{i=1}^{n} \sum_{k=1}^{c} \mu_{ik}^{\beta} \sum_{l=1}^{p} \sum_{t, a_l^{(t)} \neq \mathbf{x}_{il}} w_{kl}^{(t)}. \end{aligned}$$

From (2.29) we have

$$\sum_{t, a_l^{(t)} \neq \mathbf{x}_{il}} w_{kl}^{(t)} = 1 - \sum_{t, a_l^{(t)} = \mathbf{x}_{il}} w_{kl}^{(t)}. \tag{3.3}$$

Since the sums on objects, clusters and attributes are independent and using (3.3) the objective function can be written as

$$J_{FC}(\mathbf{U}, \mathbf{V}) = \sum_{k=1}^{c} \sum_{l=1}^{p} \sum_{i=1}^{n} \Big[\mu_{ik}^{\beta} - \mu_{ik}^{\beta} \sum_{t, a_l^{(t)} = \mathbf{x}_{il}} w_{kl}^{(t)} \Big].$$

Minimizing J_{FC} is equivalent to minimizing

$$J_{FC}(\mu_{1k}, \mu_{2k}, \ldots, \mathbf{v}_{\mathbf{k}}) = \sum_{i=1}^{n} \mu_{ik}^{\beta} - \sum_{i=1}^{n} \mu_{ik}^{\beta} \sum_{a_l^{(t)} = \mathbf{x}_{il}} w_{kl}^{(t)} \; \forall k \in [1, c], \forall l \in [1, p].$$

Since $\mathbf{U}$ is fixed, minimizing J_{FC} is equivalent to maximizing

$$J_{FC}(\mathbf{v}_k) = \sum_{i=1}^{n} \mu_{ik}^{\beta} \sum_{t, a_l^{(t)} = \mathbf{x}_{il}} w_{kl}^{(t)},$$

under the constraints (2.28) and (2.29).

As the weights w_{kl} are independent, maximizing J_{FC} is equivalent to maximizing

$$J_{FC}(w_{kl}) = \sum_{i=1}^{n} \sum_{t, a_l^{(t)} = x_{il}} \mu_{ik}^{\beta} w_{kl}^{(t)}, \ \forall k, \forall l$$

Since (see Remark 5 for more details)

$$\sum_{i=1}^{n} \sum_{t, a_l^{(t)} = x_{il}} \mu_{ik}^{\beta} w_{kl}^{(t)} = \sum_{t=1}^{n_l} S_{kl}^{(t)} w_{kl}^{(t)} \tag{3.4}$$

then the optimization problem becomes

$$\begin{cases} \max & J_{FC}(w_{kl}) = \sum_{t=1}^{n_l} S_{il}^{(t)} w_{kl}^{(t)}, \\ s.t. & (2.28) \text{ and } (2.29) \text{ are satisfied.} \end{cases} \tag{3.5}$$

For $\mathbf{U}$ fixed, the term $S_{il}^{(t)}$ is constant, the problem (3.5) then corresponds to a linear optimization problem that is solved by giving the maximal weight to the categorical value that is the most frequent in the cluster, i.e, when Equation (3.2) is satisfied. □

With the new updating of the centers, we proposed an extension of *FC* referred to as *FC**.

Remark 5. To better understand Equation (3.4), let's consider the dataset in Table 3.1 to illustrate the equation. Let say we want to compute the left term for the category *yes* of F. We can sum over objects, i.e., i to search where the category appear: $i \in \{1, 3, 4\}$. Consequently, for the category *yes*, we will consider the quantity $\mu_{ik}^{\beta} w_{kl}^{(t)}$ for $i \in \{1, 3, 4\}$. Similarly, instead of summing over i, we can sum over the categories to search the occurrences of a given category. For instance, the category *yes* appear for $i \in \{1, 3, 4\}$. The latter case corresponds to the right term of Equation (3.4).

It can be noted that the new update of $w_{kl}^{(t)}$ gives most of the time binary values. Indeed, in practice the case $S_{kl}^{(r)} = S_{kl}^{(\tau_1)} = \ldots = S_{kl}^{(\tau_{q-1})}$ is very unlikely to appear. Consequently, the algorithm generates mostly hard centers instead of fuzzy as proposed in [30]. It can also be noted that in practice *FC** is similar to *FKM*. Indeed, in practice, the two algorithms have the same updates of the centers (see Equation (2.26)) and the partition matrix.

	F
x_1	yes
x_2	no
x_3	yes
x_4	yes

Table 3.1: Illustration of Equation (3.4)

Algorithm and complexity analysis

With Equation (3.2), the algorithm of FC* can be derived from FC's in Section 2.2.8. The corresponding algorithm is summarized in **Algorithm 7**. For the time complexity, to compute the distance, $n\mathcal{J}$ operations are needed. To compute $S_{kl}^{(t)}$, n operations are needed, consequently, the time complexity to update the centers is $O(nc\mathcal{J})$. The overall time complexity is given by $O(T(n\mathcal{J}+nc+nc\mathcal{J})) \rightarrow O(nc\mathcal{J}T)$. Indeed, the time complexity of $\mathbf{U}$ is the same as *FCM*'s, i.e., $O(cn)$ with [58]. For the memory complexity, to store $\mathbf{V}$, $c\mathcal{J}$ memory size is needed. Hence, the overall memory complexity is given by $O(np + nc + c\mathcal{J})$.

Remark 6. In application, as the *FC** algorithm is similar to *FKM*'s, the time and memory complexity can be reduce respectively to $O(ncp)$ and $O(np + nc + cp)$.

Algorithm 7 *Hard centers updates of FC* algorithm (FC*)

Require: $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_i}, \ldots, \mathbf{x_n}\}$ the categorical data, $2 \leq c < n$ the number of clusters, and ϵ a stop criteria.

Output: $\mathbf{U}$ fuzzy c-partitions of $\mathbf{X}$ and $\mathbf{V}$ the centers of clusters.

Begin

Randomly initialize $\mathbf{V}_0$.

$\tau \leftarrow 0$

repeat

$\tau \leftarrow \tau + 1$

Compute the distance matrix D with the generalized Hamming distance.

Update $\mathbf{U}_\tau = [\mu_{\mathbf{ik}}]_\tau$ with Equation (2.14) where $d_{ik}^2 = d_{ik}^{GH}$.

Update $\mathbf{V}_\tau$ Equation (3.2).

until $\|J_{FC*}^{\tau-1} - J_{FC*}^{\tau}\| \leq \varepsilon$

End

To follow the original idea in [30] (obtaining fuzzy centers), we used Shannon's entropy to regularize the attributes category weights. This procedure is presented in the next section.

3.2.1 Entropy as regularization function

In subsection 2.2.4, we present some applications of Shannon's entropy in clustering. Among the presented methods, some use entropy as a regularization function. In the following paragraphs, we describe how Shannon's entropy can be used to penalize attribute category weights.

From Equation (1.5), the Shannon entropy associated to the attribute categories is given by the following equation

$$H_{n_l}(w_{kl}^{(1)}, \ldots, w_{kl}^{(n_l)}) = -\sum_{k=1}^{c}\sum_{l=1}^{p}\sum_{t=1}^{n_l} w_{kl}^{(t)} \ln(w_{kl}^{(t)}). \tag{3.6}$$

Equation (3.6) makes sense as $w_{kl}^{(t)}$ satisfies $\sum_{t=1}^{n_l} w_{kl}^{(t)} = 1$. As discussed in 1.2, H_{n_l} reaches its maximum when all the weights are equal

$$w_{kl}^{(t)} = \frac{1}{n_l}, \forall t \in \{1, \ldots, n_l\}. \tag{3.7}$$

Similarly, the minimum is reached when the weights are binary, hence

$$w_{kl}^{(t)} \in \{0, 1\} \forall t \in \{1, \ldots, n_l\}. \tag{3.8}$$

Therefore we have

$$0 \leq H_{n_l} \leq \frac{1}{n_l}. \tag{3.9}$$

In hard centers clustering such as *k-modes* and *FKM* algorithms, the weights associated to the attributes categories are binary. Applying Shannon entropy to these weights leads to a minimum entropy, i.e., a minimum disorder. Maximizing the entropy will constrain the weights to be nonbinary. However, a full maximization of the entropy leads to uniform weights. To allow soft and accurate weights, a trade-off between the maximization of the entropy and the minimization of the objective function of FC* can be defined by associating a positive coefficient to the entropy that will indicate the importance given to it.

In the following subsection, we describe the objective function of *CFE* by considering Shannon's entropy.

3.3 *CFE*: fuzzy entropy c-means

We integrated the entropy H_{n_l} into the cost function of *FC* as in [61, 62] and proposed a new clustering method for categorical data called *categorical fuzzy entropy c-means* (*CFE*). The objective function, updating formula of the partition matrix, and the centers, the algorithm and complexity analysis of *CFE* are provided in the following subsections.

Objective function

The cost function of *CFE* is given as follows

$$\begin{aligned} J_{CFE}(\mathbf{U},\mathbf{V}) &= J_{FC} - \Psi H_{n_l}(w_{kl}^{(1)},\ldots,w_{kl}^{(n_l)}) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{c}\mu_{ik}^{\beta} d_{ik}^{GH} + \Psi \sum_{k=1}^{c}\sum_{l=1}^{p}\sum_{t=1}^{n_l} w_{kl}^{(t)} \ln(w_{kl}^{(t)}), \end{aligned}$$

where $\Psi = \Phi\, n$ with Φ an input parameter that controls the importance given to the entropy and n the number of objects. The factor n in Ψ is due to the fact that we noticed in the experiments that optimal values of Ψ depend on n.

The optimization problem associated to *CFE* is then given by

$$\begin{cases} J_{CFE}(\mathbf{U},\mathbf{V}) = & \displaystyle\sum_{i=1}^{n}\sum_{k=1}^{c}\mu_{ik}^{\beta} d_{ik}^{GH} + \Psi \sum_{k=1}^{c}\sum_{l=1}^{p}\sum_{t=1}^{n_l} w_{kl}^{(t)} \ln(w_{kl}^{(t)}), \\ s.t. & \text{Equation (2.8), (2.28) and (2.29) are satisfied.} \end{cases} \tag{3.10}$$

Optimization

As in previous algorithms, the problem (3.10) can be solved using the alternate optimization scheme. When the centers $\mathbf{V}$ are fixed, the partition matrix membership degrees μ_{ik} are updated with Equation (2.14). Indeed, as the centers $\mathbf{V}$ are fixed, the distance which is a function of $w_{kl}^{(t)}$ is fixed and by applying the Lagrangian (let say $\mathcal{L}$), all the derivatives of $\mathcal{L}$ with respect to $w_{kl}^{(t)}$ $(\dfrac{\partial \mathcal{L}}{\partial w_{kl}^{(t)}})$ equal zero.

When $\mathbf{U}$ is fixed the updating of the centers is given by the following theorem.

Theorem 3.3.1. For $\mathbf{U}$ fixed, the cluster centers $\mathbf{V}$ are minimized iff

$$w_{kl}^{(s)} = \frac{\exp\left[-\frac{1}{\Psi}\sum_{\mathbf{x_{il}}\neq a_l^{(s)}}\mu_{ik}^{\beta}\right]}{\sum_{t=1}^{n_l}\exp\left[-\frac{1}{\Psi}\sum_{\mathbf{x_{il}}\neq a_l^{(t)}}\mu_{ik}^{\beta}\right]}, \quad \forall k \in \{1,\ldots,c\}, \forall l\{1,\ldots,p\}, \forall s \in \{1,\ldots,n_l\}. \tag{3.11}$$

Proof. We have

$$J_{CFE}(\mathbf{U},\mathbf{V}) = \sum_{i=1}^{n}\sum_{k=1}^{c}\mu_{ik}^{\beta} d_{ik}^{GH} + \Psi \sum_{k=1}^{c}\sum_{l=1}^{p}\sum_{t=1}^{n_l} w_{kl}^{(t)} \ln(w_{kl}^{(t)})$$

Since $\mathbf{U}$ is fixed, and the sum over clusters, and attributes are independent, given $k \in [1, c]$, $l \in [1, p]$, by using Equation (3.4), minimizing J_{CFE} is equivalent to minimizing

$$\begin{cases} J_{CFE}(w_{kl}) = \sum_{t=1}^{n_l} \sum_{i, \mathbf{x}_{il} \neq a_l^{(t)}} \mu_{ik}^{\beta} w_{kl}^{(t)} + \Psi \sum_{t=1}^{n_l} w_{kl}^{(t)} \ln(w_{kl}^{(t)}), \\ s.t. (2.28) \text{ and } (2.29) \text{ are satisfied.} \end{cases} , \forall k, \ \forall l.$$

Let $\mathcal{L} = J_{CFE}(w_{kl}) + \lambda_{kl}(\sum_{t=1}^{n_l} w_{kl}^{(t)} - 1)$ be the Lagrangian associated to the optimization problem and λ_{kl} the Lagrangian multipliers. By differentiating the Lagrangian with respect to $w_{kl}^{(s)}$ and λ_{kl} we obtain

$$\frac{\partial \mathcal{L}}{\partial w_{kl}^{(s)}} = \left[\sum_{i, \mathbf{x}_{il} \neq a_l^{(s)}} \mu_{ik}^{\beta} \right] + \Psi(1 + \ln(w_{kl}^{(s)})) + \lambda_{kl}, \tag{3.12}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_{kl}} = \sum_{t=1}^{n_l} w_{kl}^{(t)} - 1. \tag{3.13}$$

Setting Equation (3.12) to 0 gives

$$w_{kl}^{(s)} = \exp\left[-\left(1 + \frac{\lambda_{kl}}{\Psi} + \frac{1}{\Psi} \sum_{i, \mathbf{x}_{il} \neq a_l^{(s)}} \mu_{ik}^{\beta}\right)\right]. \tag{3.14}$$

Replacing $w_{kl}^{(s)}$ by Equation (3.14) in (3.13) set to 0 gives

$$\exp\left[-\left(1 + \frac{\lambda_{kl}}{\Psi}\right)\right] = \frac{1}{\sum_{t=1}^{n_l} \exp\left[-\frac{1}{\Psi} \sum_{i, \mathbf{x}_{il} \neq a_l^{(t)}} \mu_{ik}^{\beta} \right]}. \tag{3.15}$$

Reporting $\exp\left[-\left(1 + \frac{\lambda_{kl}}{\Psi}\right)\right]$ into Equation (3.14) gives

$$w_{kl}^{(s)} = \frac{\exp\left[-\frac{1}{\Psi} \sum_{i, \mathbf{x}_{il} \neq a_l^{(s)}} \mu_{ik}^{\beta} \right]}{\sum_{t=1}^{n_l} \exp\left[-\frac{1}{\Psi} \sum_{i, \mathbf{x}_{il} \neq a_l^{(t)}} \mu_{ik}^{\beta} \right]}. \tag{3.16}$$

Therefore J_{CFE} is minimized iff $w_{kl}^{(t)}$ satisfies Equation (3.11). □

Equation (3.11) shows that the application of Shannon's entropy as a regularization function of the category's weights helps to construct fuzzy centers. The new weights $w_{kl}^{(t)}$ which are now non-binary indicate the importance of each category in the clusters.

Remark 7. It can be noticed that (3.11) satisfies the constraint (2.28).

Algorithm and complexity analysis

The algorithm of *CFE* is summarized in **Algorithm 8**.

Given the number of clusters c, a chosen value of β and Φ, the first step consists in initializing the centers such that Equations (2.28) and (2.29) are satisfied. Then the cluster membership degrees μ_{ik} and the prototypes are updated using respectively Equations (2.14) and (3.11). The preceding step is repeated until there exists almost no change from an iteration to another (i.e., when $\|\mathbf{J}_{CFE}^{\tau} - \mathbf{J}_{CFE}^{\tau-1}\|$ reaches a variable ε set to a small value).

To update the weight of one attribute category, the denominator of Equation (3.11) requires $n_l n$ operations. By considering $n_l << n$, n operations are performed once for each attribute. For $\mathcal{J}$ categories, c clusters, $nc\mathcal{J}$ operations are then needed. By taking into account the time complexity to compute D and $\mathbf{U}$ with [58] of the FC algorithm (same as *CFE*), the overall time complexity is given by $O(T(nc\mathcal{J} + cn)) \rightarrow O(nc\mathcal{J}T)$. Finally, the memory complexity of *CFE* remains the same as *FC*, i.e., $O(np + nc + c\mathcal{J})$.

Algorithm 8 Categorical fuzzy entropy c-means algorithm (CFE).

Require: $\mathbf{X} = \{x_1, \ldots, x_n\}$ the categorical data, $1 < c < n$ the number of clusters, $\Phi > 0$ the fuzzy entropy weighting coefficient, $\beta > 1$ a weighting exponent, and $\epsilon \geq 0$ a threshold.
Output: $\mathbf{U}$ fuzzy c-partitions of $\mathbf{X}$ and $\mathbf{V}$ the clusters centers.
Begin
  Randomly initialize $\mathbf{V}_0$ according to Equation (2.28) and (2.29).
  $\tau \leftarrow 0$
  repeat
    $\tau \leftarrow \tau + 1$
    Compute the distance matrix D with the generalized Hamming distance.
    Update $\mathbf{U}_\tau$ using (2.14) with $d_{ik}^2 = d_{ik}^{GH}$.
    Update centers $\mathbf{V}_\tau$ using (3.11).
  until $\|\mathbf{J}_{CFE}^{\tau-1} - \mathbf{J}_{CFE}^{\tau}\| \leq \varepsilon$
End

3.4 Experiments

In this section, we conduct several experiments to test the performance of the new clustering algorithm *CFE*. We describe the datasets used, the evaluation criteria, and the experimental protocol. Finally, we demonstrate some properties (i.e., strengths) of *CFE*.

3.4.1 Datasets

We simulate two categorical datasets to illustrate the behavior of *CFE* and select nine clustering benchmark categorical datasets from the UCI repository which are characterized by different dimensions, the number of objects, and attributes categories to evaluate the performance of *CFE*. We then compare *CFE* with existing numerical and categorical clustering methods. The datasets are Zoo, Soybean, Congressional voting records, Breast Cancer, Lung, Cars, Mushrooms, Credits, Dermatology. The description of each dataset is provided in the following subsections.

Simulated data

We simulate datasets to illustrate the behavior of *CFE* in capturing fuzzy centers. As discussed in Section 2.2.8, these centers can help to interpret clusters according to the weights associated with each attribute category. To better show it, we generated datasets of 200 objects which are assigned in one of the two clusters. Each object is described by 2 attributes with two categories per attribute. We vary the probability of occurrence of categories in each cluster and compared them to the weights obtained by *CFE* with the frequency of each category.

Benchmark datasets

The benchmark datasets described below will are used to test the performance of *CFE* and to compare the latter with other clustering algorithms.

Zoo

The Zoo dataset contains 100 animals grouped in the 7 following classes reptiles, invertebrates, birds, amphibians, fishes, mammals, and insects. The number of objects per class is respectively 41, 20, 5, 13, 4, 8, 10. Each object of the dataset is described by 16 Boolean-valued attributes (hair, eggs, milk, tail, etc...) indicating whether the animals are concerned with each attribute. For instance, a value of 1 indicates that the animal has hair, and 0 specify that it hasn't.

Soybean

The soybean dataset is a sample of 46 soybean crops diagnosed with 4 diseases: Diaporthe stem canker, Charcoal rot, Rhizoctonia root rot, and Phytophthora rot. The size of the classes (diseases) is respectively given by 9, 10, 10, 17. The crops are described by 35 categorical attributes from which we can cite the date (April to October), the precipitation, and the temperature taking values in (less than normal, normal, and greater than normal).

We dropped all the attributes (14) having just one category. The number of categories for the remaining attributes varies from 2 to 7.

Congressional voting records

The congressional voting records dataset denoted by votes includes Democrats and Republicans votes for each of the United States House of representatives congressmen. The sample size described by 16 Boolean attributes is 434. The number of objects in the Democrats and Republicans classes is respectively 267 and 167.

Breast Cancer

The breast cancer dataset is a medical dataset of breast cancer diagnosis of a sample of 699 patients. Each patient is characterized by 9 attributes describing their tumor tissue. The sample is grouped into 2 classes indicating the malignant (cancerous cells) or benign (non-cancerous cells) tumors. Among the patients, 241 were diagnosed with the malignant tumor and 458 with the benign. Most of the attributes (8 over 9) have 9 categories and one attribute has 10.

Lung

The Lung dataset is a sample of 32 patients diagnosed with 3 types of pathological lung cancer (Type A, Type B, Type C). Each patient is characterized by 56 attributes extracted from the clinical data and X-ray data. The number of attributes categories varies between 2 and 3 and the number of objects per class is 9, 10, and 13 respectively for Type A, Type B, and Type C.

Cars

The Cars dataset contains a sample of 1728 cars described by 6 attributes the overall and maintenance prices (very high, high, medium, low), the number of doors (2, 3, 4, 5 or more), the capacity in terms of persons to carry (2, 4, more than 4), the size of luggage boot (small, medium, big) and the estimated safety of the cars (low, medium, high). The sample is grouped into 4 classes indicating the comfort of the car which are bad (1210 objects), acceptable (384 objects), good (65 objects), and very good (69 objects).

Mushrooms

The Mushrooms dataset contains data on the classification of mushrooms as poisonous (4208 samples) or edible (3916 samples). The data is described by 21 attributes which have between 2 and 12 categories. We dropped one attribute having a single category.

	# Objects	# Attributes	Max # of categories	# Classes
Lung	32	56	3	3
Soybean	46	21	7	4
Zoo	100	16	2	7
Breast Cancer	699	9	10	2
Dermatology	366	34	6	4
Votes	434	16	2	2
Credits	689	15	15	2
Cars	1728	6	4	4
Mushrooms	8124	21	12	2

Table 3.2: Categorical benchmark datasets.

Credits

The Credits dataset is a sample of 689 credit card application approval. Each object of the dataset is described by 15 attributes having a number of categories between 2 and 15. We dropped all the continuous attributes (6) from the dataset. From the data, the number of approved and non-approved credit cards is respectively 383 and 306.

Dermatology

The dermatology dataset is a medical database containing a sample of 366 patients diagnosed with erythema (redness of the skin caused by injury, infection, or inflammation). The data contain 34 attributes of which 12 are clinical features and 22 histopathological features determined by an analysis of the samples under a microscope. Every feature except the family history and the age was given a degree in the range of 0 to 3 where 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicates the relative intermediate values. The family history feature has the value of 1 if any of the 7 erythema diseases was observed in the family of the patient and 0 otherwise. The sample size of each disease annotated from 1 to 6 are respectively 112, 72, 61, 52, 49, 20.

A summary of the characteristics of datasets is provided in Table 3.2. To visualize datasets, we performed a dimensionality reduction with principal components analysis (PCA) by converting datasets to numerical with the one-hot encoding technique. The first two dimensions are plotted in Figure 3.1. In this figure, colors correspond to the real classes of objects. In the next subsection, we describe the evaluation criteria used to compare the clustering algorithms.

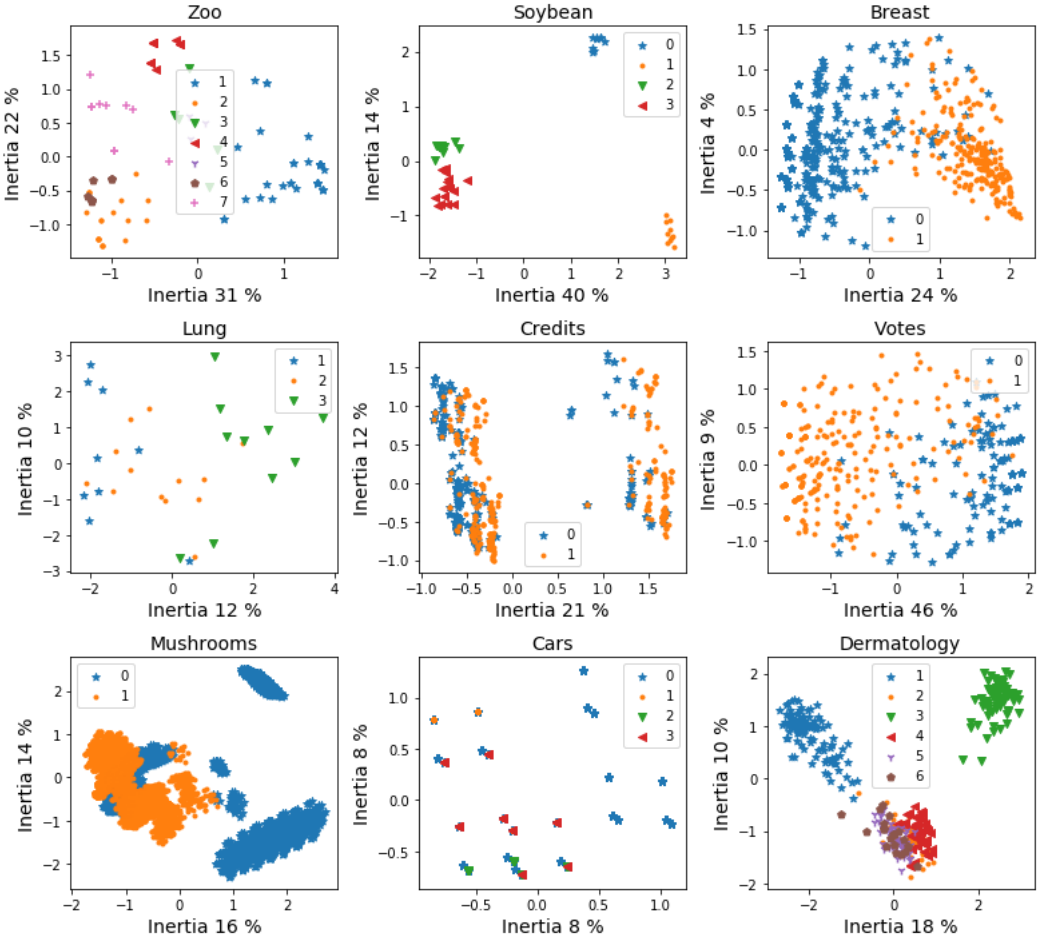

Figure 3.1: PCA results on datasets. The inertia on the axes indicates the amount of explained variance.

3.4.2 Evaluation criteria

To measure the performance of a clustering method, two main evaluation criteria can be used: internal and external criteria. Internal evaluation criteria measures describe the goodness of fit of a method from the partition matrix. In contrast, for external measures, real object classes are needed to evaluate the similarity between the latter with the produced partitions of methods. It should be noted that, in general, internal measures are the most used evaluation criteria as the real classes are not always known.

In the following subsections, we provide the descriptions of internal and external evaluation criteria used in the comparisons.

Internal measures

In the literature, several fuzzy clustering internal validity measures have been proposed (see [69] for a review). Among them, Bezdek proposed the partition coefficient, and the partition entropy described as follows.

The partition coefficient index (PC) [26] corresponds to the average quadratic sum of the fuzzy membership degrees μ_{ij}. It indicates the relative amount of membership sharing between pairs of fuzzy subsets in $\mathbf{U}$. The index is given by the following equation

$$PC = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c}\mu_{ik}^2. \tag{3.17}$$

The PC varies in the interval $[1/c, 1]$ and can be used for selecting the optimal number of clusters by maximizing Equation (3.17).

In [26, 70, 71], Bezdek introduced the partition entropy (PE) validity index which measures the amount of fuzziness in a given partition $\mathbf{U}$. It is given by the following equation

$$PE = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c}\mu_{ik}\log(\mu_{ik}). \tag{3.18}$$

The PE index is bounded by 0 and $\ln(c)$ and can be used as the PC index to determine the optimal number of clusters by minimizing Equation (3.18).

In addition to the PC and PE indexes, we use the fuzzy silhouette index FS [72], an extension of the silhouette index [73] to the fuzzy framework to quantify the goodness of separations of clusters. It is defined by

$$FS = \frac{\sum_{i=1}^{n}(\mu_{pi} - \mu_{qi})^{\theta} s_i}{\sum_{i=1}^{n}(\mu_{pi} - \mu_{qi})^{\theta}}, \tag{3.19}$$

where μ_{pi} and μ_{qi} are the first and second largest membership degrees of i_{th} object, $\theta \geq 0$ a weighting exponent and s_i is the silhouette index defined by

$$s_i = \frac{b_i - a_i}{max(b_i, a_i)}, \tag{3.20}$$

where a_i is the average distance of between the i_{th} object and all objects in the same cluster, and b_i the minimum distance between i_{th} object to its nearest cluster it is not a part of.

As the silhouette index obtained when $\theta = 0$, the FS index varies from -1 to 1, with 1 indicating an optimal value. In all our experiments, we set the value of θ to 1.

External measures

We use the Rand index (RI) [74] as an external measure to compare the predicted and real classes of objects. The RI computes the similarity measure between the predicted and real classes by considering all pairs of samples and counting pairs allocated to the same or separate clusters in the predicted and real classes.

Let a and b be the pairs of samples assigned in respectively the same and different clusters. The RI is defined by

$$RI = \frac{2(a+b)}{n(n-1)}. \tag{3.21}$$

The RI score varies between 0 and 1 with higher values corresponding to a good matching of the predicted and real classes of objects. A value of RI close to 0.5 indicates a random labeling of objects.

As a generalization of the RI, we also use the fuzzy rand index (FRI) [75] to compare the fuzzy partitions denoted by $Q = [q_{ij}]_{n\times c}$ obtained from the compared methods to the real objects classes (reference hard partition) denoted by $R = [r_{ik}]_{n\times c}$. Contrary to the RI, the FRI takes advantage of the fuzzy assignment of the objects to the clusters hence, it captures more information on the similarity between the obtained fuzzy partitions and the true classes. Moreover, the FRI has the ability to compare different fuzzy partitions among them, i.e, the reference partition can be a fuzzy partition obtained from the compared methods.

Let V, X, Y, and Z be four fuzzy sets such that:

- V is the fuzzy set of pairs of objects belonging to the same class in R.

- X is the fuzzy set of pairs of objects belonging to different classes in R.
- Y is the fuzzy set of pairs of objects belonging to the same cluster in Q.
- Z is the fuzzy set of pairs of objects belonging to different clusters in Q.

Let (i_1, i_2) be a pair of different objects (in R or in Q), k_1 and k_2 be two different classes in R and j_1 and j_2 be two different clusters in Q. Let $\sqsupset$ be a triangular norm (e.g. min) and $\urcorner$ a triangular co-norm (e.g. max). In the following equations, $g \sqsupset h$ and $g \urcorner h$ denotes respectively the triangular norm and co-norm between g and h.

The FRI is defined in [75] by

$$FRI = \frac{|V \cap Y| + |X \cap Z|}{|V \cap Y| + |V \cap Z| + |X \cap Y| + |X \cap Z|}, \tag{3.22}$$

where

$$|V \cap Y| = \sum_{i_1=1}^{i_2-1} \sum_{i_2=2}^{n} V(i_1, i_2) \sqsupset Y(i_1, i_2) \tag{3.23}$$

$$|X \cap Z| = \sum_{i_1=1}^{i_2-1} \sum_{i_2=2}^{n} X(i_1, i_2) \sqsupset Z(i_1, i_2) \tag{3.24}$$

$$|V \cap Z| = \sum_{i_1=1}^{i_2-1} \sum_{i_2=2}^{n} V(i_1, i_2) \sqsupset Z(i_1, i_2) \tag{3.25}$$

$$|X \cap Y| = \sum_{i_1=1}^{i_2-1} \sum_{i_2=2}^{n} X(i_1, i_2) \sqsupset Y(i_1, i_2) \tag{3.26}$$

with

$$\begin{aligned}
V(i_1, i_2) &= (r_{i_1 1} \sqsupset r_{i_2 1}) \urcorner \ldots \urcorner (r_{i_1 k} \sqsupset r_{i_2 k}) \triangleq \urcorner_{k=1}^{c} (r_{i_1 k} \sqsupset r_{i_2 k}) \\
Y(i_1, i_2) &\triangleq \urcorner_{k=1}^{c} (q_{i_1 k} \sqsupset q_{i_2 k}) \\
X(i_1, i_2) &\triangleq \urcorner_{k_1, k_2 = 1/k_1 \neq k_2}^{c} (r_{i_1 k_1} \sqsupset q_{i_2 k_2}) \\
Z(i_1, i_2) &\triangleq \urcorner_{j_1, j_2 = 1/j_1 \neq j_2}^{c} (q_{i_1 j_1} \sqsupset q_{i_2 j_2}))
\end{aligned}$$

We use the presented five measures to evaluate and compare the performance of *CFE* to existing methods. The next subsection describes how the evaluations and comparisons are performed.

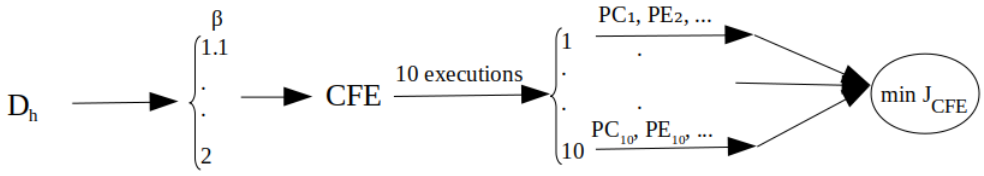

Figure 3.2: Experimental protocol where D_h are datasets with $h = 1, \ldots, 9$.

3.4.3 Experimental protocol

To test the performance of the *CFE*, we perform several experiments. We run the algorithm on the nine datasets described in Section 3.4.1. For each dataset, we vary the value of the parameter β between 1.1 and 2 and execute the algorithm for each value ten times with different initializations. Then, we select the scores PC, PE, FS, RI, and FRI for the execution with the lowest cost J_{CFE}. This protocol is summarized in Figure 3.2.

Remark 8. The choice of the scores at the lowest cost is due to the fact the k-means-based clustering algorithms are very sensitive to the initialization [76]. The lowest cost among the considered iterations corresponds to the best local minimum of J_{CFE} and overcomes the variability of the scores from one iteration to another.

We compare the performance of *CFE* and *FC**[1] based on the five preceding scores with the *k-modes*, and *FCM* algorithms over the nine datasets. For *FCM*, we use the one-hot encoding to transform datasets to numerical.

We follow the recommendations of Demsar in [77] to compare the methods over the nine datasets. The recommendations in [77] can be described as follows: given the scores (e.g. RI) over the considered datasets, a Friedman test [78, 79] at a significance level of γ (e.g. 0.05) is firstly performed to rank the algorithms with the best performing algorithm having rank the first rank. If the test is significant, a Wilcoxon signed ranks test [80] is performed for pairwise comparisons of the models. In [77], Demsar introduced a graphical visualization called critical difference diagram to represent the results of the statistical analysis. In the latter diagram, non-statistically significant algorithms are connected with a bold line while statistically significant algorithms are not connected.

3.4.4 Parameter settings

For all the compared methods we vary the value of the parameter β from 1.1 to 2 as described in Figure 3.2 and we set the number of clusters c for

[1] This method is similar to *FKM* in practice.

each dataset to the known values in Table 3.2.

The value of the parameter Ψ is critical in *CFE*. If it is too high, the entropy in the cost function will be considered more important. Therefore, the attributes category weights will be highly penalized and in some situations, all the weights will reach the same value $1/n_l$. If the value of Ψ is too low, more importance will be accorded to J_{FC}. In this situation, only a small amount of weights will be greater than 0. Consequently, *CFE* will behave as a crisp centers algorithm.

As n in Ψ is fixed, and $\Psi = \Phi n$, we conduct prior experiments on the nine datasets in which we vary the value Φ. The optimal value in term of a good balance between J_{FC} in the cost function of *CFE*, and the entropy is $\Phi = 0.01$. The same value provides good performances based on the scores on all datasets.

In the next subsection, we describe the materials in the experiments.

3.4.5 Materials

To conduct the experiments we use several external software such as the Python packages kmodes [81] (version 0.11.0) and scikit-fuzzy [82] (version 0.4.2) respectively for *k-modes* and *FCM* algorithms. For the scores FRI and FS, we use the R package fclust [83] (version (2.1.1)). We use the code provided in [84] for the critical difference diagrams.

We provide the code source of the implementations of *CFE* and *FC** on Github[2].

In the next subsection, we describe the results of the experiments.

3.4.6 Results

In this subsection, we first discuss the results of *CFE* obtained on the nine datasets, then we present the results of the statistical comparisons between *CFE*, *FC**, *k-modes*, and *FCM*.

CFE results

Tables A.1.1, A.1.2, A.1.3, A.1.4 and A.1.5 from Appendix A respectively describe the scores of RI, FRI, FS, PC, and PE obtained with *CFE* on the nine benchmark datasets. It can be noted that, for all scores, the values differ from a dataset to another. For the Credits, Votes, and Cars datasets

[2] https//github.com/abdjiber/cfe

the scores of RI, FRI, and FS are approximately the same for all values of β. For all datasets, the FRI scores are lower than RI's which can be interpreted by the fact that FRI scores behave like an adjustment of the RI scores by taking into account the fuzziness of the partitions.

Despite the low scores of RI, FRI, and FS of *CFE* on some datasets (e.g. Cars and Mushrooms) the corresponding PC scores are high which can be interpreted by the fact that *CFE* captures the fuzziness in these datasets. Similarly, it can be noted that for all datasets, optimal scores of RI and FRI are obtained with low values of β whereas, for the FS scores, optimal values are obtained with high values of β.

Table A.1.4 shows that as the value of β increases, the PC scores decrease (inversely for the PE scores). The latter observation is expected as β controls the fuzziness of the partitions. Indeed, small (respectively high) values of β will lead to crisp-like (respectively fuzzy-like) partitions. Therefore, optimal values of PE and PC are expected when the value of β is small respectively high.

The analysis of the scores in the preceding paragraph shows that based on the scores used, a trade-off should be made on the choice of the value of the parameter β. Either choosing to have crisp-like partitions of *CFE* which give optimal scores of PE and PC but also RI, FRI or to have fuzzy-like partitions and more separable clusters which give optimal scores of FS. From the later remarks, in order to take advantage of the ground truth classes of objects, in the remaining experiments, we set the value of β to 1.1.

In the following subsection, we present the results of the statistical analysis.

Statistical comparisons results

Remark 9. As the partitions of the *k-modes* algorithm are crisp, we compare the latter to *CFE*, *FC** and *FCM* only based on the RI scores. It should be noted that the PC and PE are at their optimal values (1 respectively 0) for any datasets for *k-modes* algorithm.

In the statistical analysis, we firstly set the significance level γ value to 0.05. For this value, the difference of performances of all the models with all scores, values of β, and overall datasets was not statistically significant. We then increase the value of γ to 0.1. For this new value, we observe a difference of performance for some values of β for the PC and PE scores.

For the PC scores, the significance is observed for $\beta \in \{1.5, 1.6, 1.7, 1.8, 1.9\}$ whereas for the PE scores, it is observed for $\beta \in \{1.5, 1.6, 1.7\}$. For the FRI

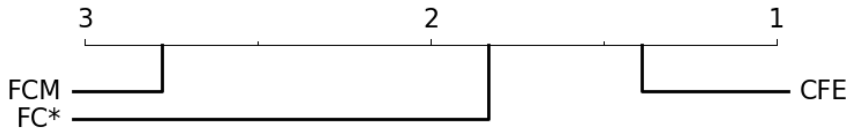

Figure 3.3: Critical difference diagram obtained for PC scores with $\gamma = 0.1$ and $\beta = 1.5$.

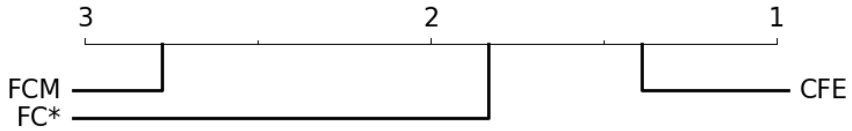

Figure 3.4: Critical difference diagram obtained for PE scores with $\gamma = 0.1$ and $\beta = 1.5$.

scores, the significance is observed only for $\beta = 1.9$ and $\gamma = 0.2$. For the RI scores, the significance is observed only for $\beta = 2$ and $\gamma = 0.25$. Finally, for the FS scores, the significance is observed only for $\gamma = 0.7$ and $\beta \in \{1.2, 1.8\}$.

Remark 10. Among the values of β and γ for which the statistical tests were significant for the scores PE, PC, and FS, in the following paragraph, we present only one critical difference diagram. The remaining diagrams are available in section A.2 from Appendix A.

Figures 3.3, 3.4, 3.5, 3.6 and 3.7 correspond respectively the critical difference diagrams obtained for PC, PE, FRI, RI and FS scores where the Friedman test is significant respectively for $\beta = 1.5$, $\beta = 1.5$, $\beta = 1.9$, $\beta = 2$, $\beta = 1.2$.

In Figure 3.3, the difference of PC scores between all the compared models is significant (all the models are not connected). From the critical difference diagram, *CFE*, *FC** and *FCM* are respectively ranked first, second and third. Therefore, with $\beta = 1.5$, *CFE* has overall the best PC scores over all datasets.

In Figure 3.4 similarly to the PC scores and $\beta = 1.5$, *CFE* outperform *FC** and *FCM* overall on all datasets.

In Figure 3.5, while the difference of FRI on all datasets is significant between (*CFE*, *FCM*) and (*FC**, *FCM*), it is not significant between *FC** and *CFE*. However, *CFE* has the first rank.

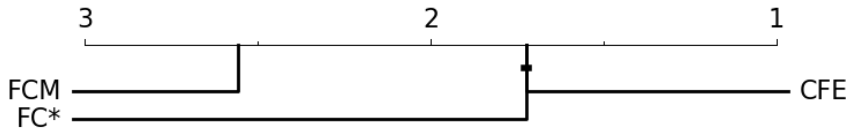

Figure 3.5: Critical difference diagram obtained for FRI scores with γ = 0.2 and $\beta = 1.9$.

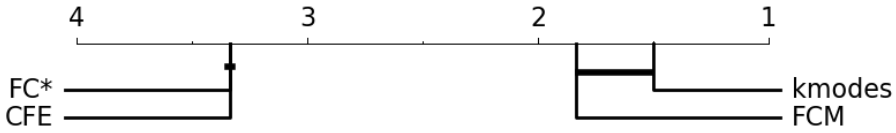

Figure 3.6: Critical difference diagram obtained for RI scores with γ = 0.25 and $\beta = 2$.

In Figure 3.6, the difference of RI scores is significant between (*k-modes*, *FCM*) and (*CFE*, *FC**) but not significant between the algorithms of each tuple.

The difference between the critical difference diagrams of FRI and RI can be explained as follows: in Figure 3.5, the FRI scores is significant between *CFE*, *FC**, and *FCM* only for the value of γ = 0.2 which correspond to two times the significance level of the PE and PC scores. In addition, it can be noted in Table A.6 which contains the data used for the statistical comparisons of the models on the FRI scores in Section A.2.1 from Appendix A, that *FCM* achieved lower performances compared to *CFE*, and *FC** particularly on the Soybean, Zoo and Mushrooms datasets. For the RI comparisons, contrary to the *CFE* and *FCM* methods which performances are influenced by the value of β (good performances respectively for low values of β such as in Tables A.7 for *CFE* and high values of β such as in Table A.7 for *FCM*), the *k-modes* performances are not affected by this parameter.

Finally, in Figure 3.7, the difference of FS scores over all datasets is significant between the compared models.

In the next section, we illustrate the strengths of *CFE* to capture fuzzy centers.

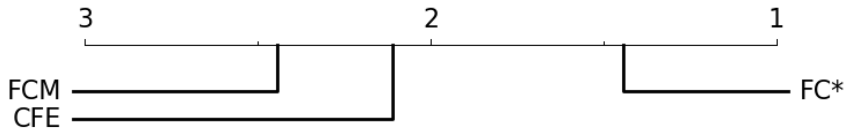

Figure 3.7: Critical difference diagram obtained for FS scores with $\gamma = 0.7$ and $\beta = 1.2$.

Attributes	Categories	C_1	C_2
A_1	0	0.8	0.27
	1	0.2	0.73
A_2	0	≈ 0	≈ 1
	1	≈ 1	≈ 0

Table 3.3: category weights obtained with *CFE* with $\Phi = 0.05$ on the simulated data for $p = 0.8$.

3.5 Strengths of *CFE*

In the following experiments, we demonstrate the behavior of *CFE* to capture fuzzy centers of clusters. To perform so, we randomly generate as in the protocol described in Section 3.4.1 categorical datasets with 2 clusters C_1 and C_2, 200 objects with 100 per cluster, 2 attributes A_1 and A_2 and 2 categories per attributes 0 and 1 such that the probability of occurrence are p and $1-p$ respectively for categories 0 and 1 for A_1 and $1-p$ and p for A_2 in C_1. Conversely, in C_2 the probabilities are respectively $1-p$ and p for categories 0 and 1 for A_1 and p and $1-p$ for A_2. It should be noted that the generation of attributes are independent conditionally to the cluster.

For $p = 0.8$, the generated dataset is described by histograms in Figure 3.8. With this value of p, the weights of categories 0 and 1 are respectively expected to be close in C_1 to 0.8 and 0.2 for A_1 and 0.2 and 0.8 for A_2. In C_2, they are expected to be close to 0.2 and 0.8 for A_1, and 0.8 and 0.2 for A_2. We obtain the weights described in Table 3.3 with *CFE* for $\Phi = 0.05$ and $\beta = 1.1$. In the latter table, the weights of categories of A_1 in C_1 are as expected. For A_2, they are slightly different. In C_2, the weights are different from the ones expected.

We repeat the preceding experience with the same configurations of *CFE* by setting p to 0.5. The results are summarized in Table 3.4. In the new experience, despite the fuzziness of the centers, the weights are still different from the ones expected.

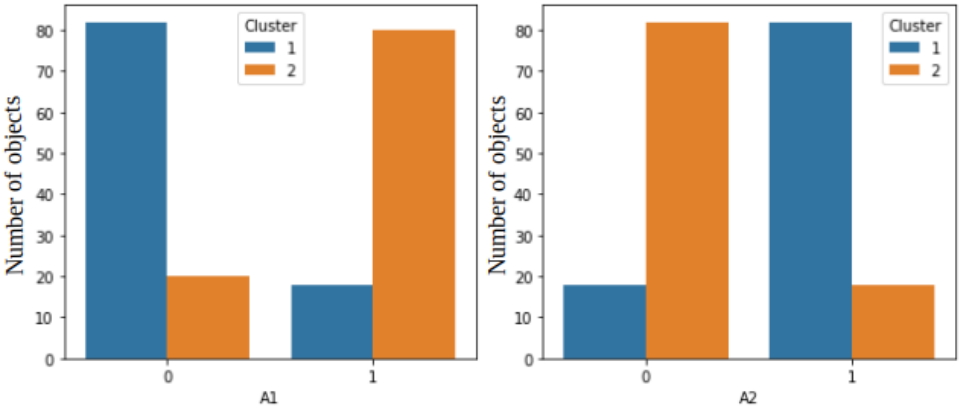

Figure 3.8: Simulated categorical data with $p = 0.8$.

Attributes	Categories	C_1	C_2
A_1	0	0.6	0.2
	1	0.4	0.8
A_2	0	≈ 0	≈ 1
	1	≈ 1	≈ 0

Table 3.4: category weights obtained with *CFE* with $\Phi = 0.05$ on the simulated data for $p = 0.5$.

In Section 3.6.2, we discuss the reasons why the weights obtained with *CFE* are different from the probabilities used in the data generation.

In the following section, we present some limitations of *CFE* and how to overcome them.

3.6 Limitations of *CFE*

As an extension of the *FCM* algorithm, *CFE* inherits most of the limitations of the latter. In this section, we discuss the limitations on choosing optimal values of the input parameters β and Φ, the non-correspondences between categories frequencies and the weights obtained with *CFE*. Finally, we discuss the difficulty to interpret the fuzzy membership values u_{ij}.

3.6.1 Optimal values of β and Ψ

It is known in the literature that the value of the input parameter β called fuzzier or fuzziness parameter has an impact on the performance and results of FCM-like clustering algorithms (see for instance [85, 86]). For numerical

data, it is generally suggested to set the value of β in $[1.5, 2.5]$ [87]. Despite the propositions by several authors in the literature of solutions to set optimal values for β, there is no formal consensus.

As a consequence, the preceding experiments show that an optimal value of β for a given dataset will not necessarily be optimal for another one. It can be also noted that in these experiments, an increase of the value of β leads in general to a decrease of performance of *CFE* such as RI, FRI scores. Based on the results of the experiments, we recommend setting low values of β in *CFE* when the desired output partitions are wanted to be crisp-like and high values of β when fuzzier partitions are desired.

Similar to β, there is also no consensus for setting an optimal value of Ψ, therefore for Φ. For numerical data, the authors in [88] proposed an iterative updating of Φ described in Equation (3.27) during the minimization of an objective function similar to *CFE*.

$$\Phi(\tau) = \Phi_0 \exp(-\frac{\tau}{\pi}), \tag{3.27}$$

where Φ_0 is the initial value of Φ, τ is the iteration number and π is an input parameter.

In our theoretical work and experiments, the implementation of Davé's solution is not tested. In contrast, we conducted prior experiments with different values the value of Φ and chose the optimal one overall datasets. Consequently, we recommend the same procedure for determining an optimal value of Φ.

In the next subsection, we discuss the difference between attributes category frequencies and the weights of these categories obtained with *CFE*.

3.6.2 Category frequencies vs weights from *CFE*

The results of experiments in Section 3.5 show that the attribute category frequencies are different with the weights obtained with *CFE* in most of the cases. We discuss in this subsection the reasons behind the non-equivalences.

Let's first consider the case where the frequencies of categories are different such as in Table 3.5. In this dataset, we consider 6 objects (3 per cluster) described by 2 attributes.

	A_1	A_2
$\mathbf{x_1}$	yes	no
$\mathbf{x_2}$	no	yes
$\mathbf{x_3}$	yes	no
$\mathbf{x_4}$	no	yes

Table 3.6: $p_{yes} = p_{no} = 0.5$

	A_1	A_2	Clusters
$\mathbf{x_1}$	yes	no	C_1
$\mathbf{x_2}$	no	yes	C_1
$\mathbf{x_3}$	yes	no	C_1
$\mathbf{x_4}$	no	yes	C_2
$\mathbf{x_5}$	yes	no	C_2
$\mathbf{x_6}$	no	yes	C_2

Table 3.5: Simulated dataset to compare the attribute categories frequencies and their weights from *CFE*.

Based on the frequencies, we would expect the following cluster centers

- $\mathbf{C}_1$ $\{\approx \frac{2}{3}/yes,\ \approx \frac{1}{3}/no\}$
- C_2 $\{\approx \frac{1}{3}/yes,\ \approx \frac{2}{3}/no\}$

Let $\mathbf{U} = [\mu_{ik}]_{1\leq i\leq 6; 1\leq k\leq 2}$ be the partition coefficient matrix obtained from *CFE*.

With Equation (3.11) we have

$$w_{11}^{(yes)} = \frac{\exp\left[-\frac{1}{\Psi}\sum_{x_{i1}\neq \mathbf{yes}} \mu_{i1}^{\beta}\right]}{\exp\left[-\frac{1}{\Psi}\sum_{x_{i1}\neq \mathbf{yes}} \mu_{i1}^{\beta}\right] + \exp\left[-\frac{1}{\Psi}\sum_{x_{i1}\neq \mathbf{no}} \mu_{i1}^{\beta}\right]} \tag{3.28}$$

The same preceding calculations can be performed for categories *yes* and *no* in $\mathbf{C}_1$ and C_2. Equation (3.28) shows that in most of the cases the weights $w_{jl}^{(t)}$ are not a linear function of the frequencies of the corresponding categories, i.e., it should not be expected for instance the weights to be equal to the frequencies or two times, etc...

Let's now consider the case where the frequencies of categories are the same such as in Table 3.6. From this table, objects $\mathbf{x_1}$ and $\mathbf{x_3}$ belong to the same cluster C_1 while $\mathbf{x_2}$ and $\mathbf{x_4}$ belong to C_2. The partition matrix U from *CFE* will look like Table 3.7 with δ close to 1.

	C_1	C_2
$\mathbf{x_1}$	δ	$1-\delta$
$\mathbf{x_2}$	$1-\delta$	δ
$\mathbf{x_3}$	δ	$1-\delta$
$\mathbf{x_4}$	$1-\delta$	δ

Table 3.7: Partition matrix format from Table 3.6.

From Table 3.7 and Equation (3.11) we have

$$\begin{aligned} w_{11}^{(yes)} &= \frac{1}{1+\exp\left[-\frac{1}{\Psi}(\mu_{11}^{\beta}+\mu_{31}^{\beta}-(\mu_{21}^{\beta}+\mu_{41}^{\beta}))\right]} \\ &= \frac{1}{1+\exp\left[-\frac{2}{\Psi}(\delta^{\beta}-(1-\delta)^{\beta})\right]} \end{aligned} \tag{3.29}$$

Clearly the value of $w_{11}^{(yes)}$ from Equation (3.29) is different from 0.5 for almost all the values of δ.

The two previous experiments explain the reasons why the weights of attributes categories obtained with *CFE* are different from their frequencies. We hypothesize that this observation is first due to the distance used by *CFE* and second to the entropy as a penalization function of the weights.

3.6.3 Interpretation of fuzzy membership degrees

Despite the ability of *FCM*-based algorithms to capture imprecision and vagueness in data, the membership degrees to clusters are hard to interpret. As explained in Section 1.3.1, depending on the value of the membership, basics interpretations such, the objects belong, partially belong, or fully belong to clusters, or as similarities may not be sufficient in some applications. Indeed we may need to know for instance the degree of belief of the assignments of objects into the cluster (i.e., is it certain or uncertain that the object belongs to the clusters ?). In other words, we would want to quantify the uncertainty of the assignments of objects into clusters. This limitation is overcome with the *cat-ECM* algorithm presented in the next chapter.

Summary

In this chapter, we present the *CFE* algorithm, a new clustering method for categorical data that uses the fuzzy sets theory for object assignment in the clusters and the representation of the centers. We conduct several experiments to test the performance and the behavior of the new method. First, we compared the new method to existing numerical and categorical

clustering methods namely *k-modes*, *FC** (*FKM*), and *FCM*. The results of the comparisons based on internal and external evaluation criteria showed that in most cases, the difference of performances is not statistically different. While the Friedman test is significant for the scores PC and FE for low values of β, larger values were needed for the other scores.

Second, we demonstrate the ability of *CFE* to successfully capture fuzzy centers even if the attributes categories' weights obtained from *CFE* are different from their frequencies. We conduct two experiments in which we showed the reasons why the differences are observed.

Despite the non-significance of the performance difference, *CFE* reached the accuracy of the numerical clustering method *FCM* and contrary to the latter there is no need to encode the input data which is in some cases can lead to an explosion of the dimensions of the original data. Through the fuzzy representation of the centers, *CFE* offers a better way to determine the attributes categories that contribute the most to the centers compared to hard centers. Therefore, *CFE* can provide a better explainability of the results compared to hard centers clustering methods such as *k-modes* and *FC**.

As with many clustering methods, *CFE* has some limitations we discuss and present some solutions.

Key points

- *CFE* is an extension of the *FC* algorithm that uses Shannon's entropy as regularization of the attributes category weights to obtained fuzzy centers.
- In the experiments, 6 clustering evaluations criteria including internal and external measures were used to compared *CFE* and *FC** with existing numerical (*FCM*) and categorical clustering methods (*k-modes*). On overall, the statistical tests are not significant. Nevertheless, *CFE* achieved the with some parameters configurations the best performances.
- The experiments showed that a trade-off between obtaining crisp-like or fuzzy-like partitions from *CFE*, should be made on the choice of the value of β.
- The time and memory complexity of *FC** and *CFE* are respectively $O(nc\mathcal{J}T)$ and $O(np + nc + c\mathcal{J})$.

Publications and Communications

The *CFE* algorithm was first presented at the IEEE International Conference on Fuzzy Systems in 2020, then a french version of the paper was presented at the conference EGC 2021. In addition, the algorithm has been presented in several seminars. We provide below the main references.

- A. J. Djiberou Mahamadou, V. Antoine, E. M. Nguifo and S. Moreno, "Categorical fuzzy entropy c-means" 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, UK.
- A. J. Djiberou Mahamadou, V. Antoine, E. M. Nguifo and S. Moreno, "Apport de l'entropie pour les c-moyennes floues sur des données catégorielles", EGC 2021.
- Rencontres des jeunes chercheurs africains en France. "Apport de l'entropie pour les c-moyennes floues sur des données catégorielles", Decembre 2020.
- MINERS[3] seminars.

[3]LIMOS workgroup on data mining https://limos.fr/workgroup/miners.

4 — Categorical evidential c-means

Contents

4.1 The need for subsets of clusters 100
4.2 Fuzzy centers and distance 101
4.3 cat-ECM: categorical evidential c-means 102
4.4 Experiments 105
4.4.1 Data sets 106
4.4.2 Evaluation criteria 106
4.4.3 Experimental protocol 107
4.4.4 Parameter settings 108
4.4.5 Materials 108
4.4.6 Results 108
4.5 Strengths of cat-ECM 111
4.6 Limitations of cat-ECM 117
4.6.1 Optimal values of β and α 117
4.6.2 Time complexity 117

Introduction

This chapter presents a new clustering method for categorical data that uses the Dempster-Shafer theory to model uncertainty in the data. The new method called categorical evidential c-means and referred to as *cat-ECM* provides a more flexible way for object assignment in clusters and a better interpretation of membership degrees. We conduct several experiments on

different datasets 1) to compare the performance of the new method against *CFE* and existing methods, and 2) to illustrate the properties of *cat-ECM*. The chapter ends with a discussion on the limitations of the new method.

4.1 The need for subsets of clusters

Traditional clustering methods either numerical or categorical generally assign objects into single clusters. As discussed previously, the fuzzy sets theory when applied to clustering offers a more flexible way to assign objects into clusters compared to the hard sets theory. However, assignments into subsets of clusters in some scenarios (real or simulated) may be needed. For instance, if C_1 and C_2 are two clusters, there might be objects that belong to the subset C_{12}, meaning that they belong to one of the clusters either C_1 or C_2 but there is an uncertainty to fully assign them in the corresponding cluster. Therefore, they are assigned into the subset C_{12}.

To better understand the notion of subsets, let's consider the dataset in Figure 4.1 is an ordinal categorical version of the Butterfly dataset used in [59]. This dataset contains 13 objects and has a particular geometric shape. Indeed, while objects in $C_1 = \{o1, o2, o3, o4\}$ and objects in $C_2 = \{o8, o9, o10, o11\}$ are "close" to each other within the same group (Cluster C_1 or cluster C_2), objects o5, o6 and o7 are "between" the previous ones, i.e, these objects are expected to belong to C_{12}. Objects o12 and o13 can be seen as "far" from all objects despite the fact that objects o6 and o13 have the same coordinate on the x-axis. As these objects are "far" from all the other objects (i.e., given a distance threshold between objects and the centers of the clusters, the distance of these objects is above the threshold) they can be considered as outliers (represented by the empty set). When traditional clustering algorithms are run on this dataset, only singletons will be generated as subsets of clusters.

Remark 11. The term ordinal in Figure 4.1 is due to the fact that in this figure, an order is considered between the attribute categories which is not the case in categorical data. Therefore, this figure corresponds to one of the possible representations of the data.

More generally, we want a clustering method that given a number of clusters c, produces a partition containing 2^c subsets of clusters, and that can also handle outliers. In the literature, the *ECM* algorithm described in Section 2.2.3 was proposed for numerical data to fit this need. To our best knowledge, there exists no variant of *ECM* that can handle categorical.

To adapt the numerical *ECM* algorithm to categorical data, there are two needs: an adaptation of the centers of the clusters and the distance. In

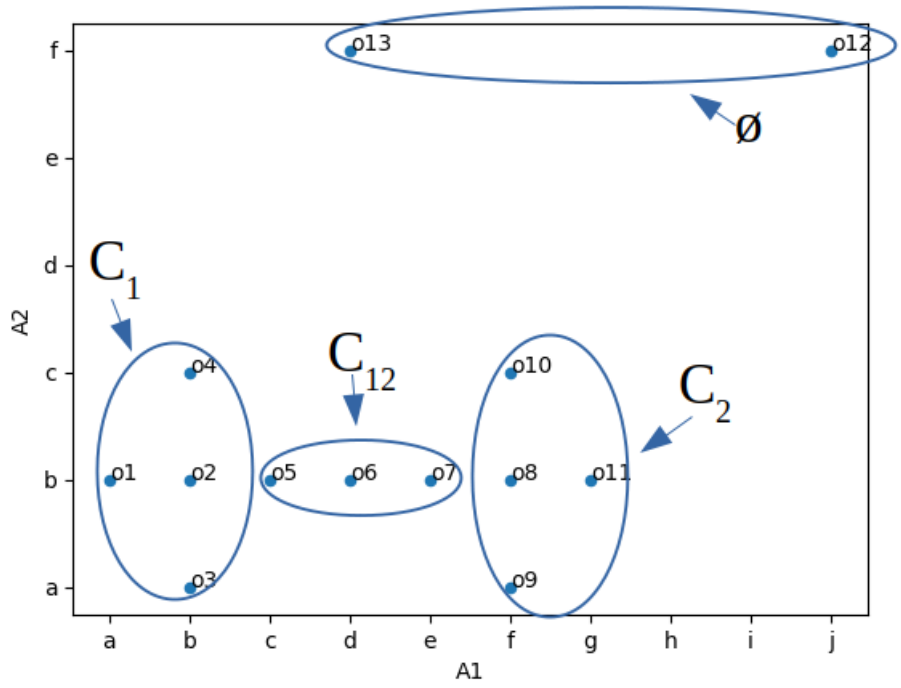

Figure 4.1: Ordinal categorical version of the Butterfly dataset.

this next section, we provide a new cluster centers definition inspired from the fuzzy centers introduced in [30] and consider the generalized Hamming distance as in *CFE*.

4.2 Fuzzy centers and distance

Let $F = (F_1, \ldots, F_l, \ldots, F_p)$ be a set of p categorical attributes and $Dom(F_l) = (a_l^{(1)} \ldots a_l^{(n_l)})$ be the domain of possible values for the attribute F_l, $\Omega = \{\omega_1, \ldots, \omega_k, \ldots, \omega_c\}$ the frame of discernment with c the number of desired clusters. Similarly to *CFE*, we define the fuzzy centers $\mathbf{v}_k = (\mathbf{v}_{k1}, \ldots, \mathbf{v}_{kl}, \ldots, \mathbf{v}_{kp})$ with $\mathbf{v}_{kl} = \{w_{kl}^{(1)}/a_l^{(1)}, \ldots, w_{kl}^{(t)}/a_l^{(t)}, \ldots, w_{kl}^{(n_l)}/a_l^{(n_l)}\}$ such that

$$0 \leq w_{kl}^{(t)} \leq 1, \tag{4.1}$$

and

$$\sum_{t=1}^{n_l} w_{kl}^{(t)} = 1 \quad \forall l \in \{1 \ldots p\}, \forall \mathbf{A}_k \subseteq \Omega, \mathbf{A}_k \neq \emptyset. \tag{4.2}$$

where

- When $|\mathbf{A}_k| = 1$, $w_{kl}^{(t)}$ are obtained through the optimization of the cost function of *cat-ECM* (see next subsection).

- When $|\mathbf{A}_k| > 1$, i.e, $c+1 \leq k \leq 2^c$,, similarly the barycenter of single clusters centers of *ECM* in [59], we propose to use the mean of attributes category weights of single clusters for subsets of clusters with cardinality greater than one, i.e,

$$w_{kl}^{(t)} = \frac{1}{|\mathbf{A}_k|} \sum_{\omega_\nu \in A_k} w_{\nu l}^{(t)} \tag{4.3}$$

To better understand this new notion of fuzzy sets, let's consider the Example 2.2.4. Let $\mathbf{v_1} = (\{0.6/yes, 0.4/no\},\ \{0.8/yes, 0.2/no\})$ be the center of C_1 and $\mathbf{v_2} = (\{0.5/yes, 0.5/no\},\ \{0.1/yes, 0.9/no\})$ the center of C_2. From Equation (4.2), the center corresponding to the subset $\Omega = \{C_1, C_2\}$ is given by

$$\mathbf{v_\Omega} = (\{0.55/yes, 0.45/no\},\ \{0.45/yes, 0.55/no\}).$$

For the distance, we consider a new variation of the generalized Hamming distance (2.31) by considering the subsets of clusters. We define the new distance as follows:
Let $\mathcal{D}_{il} = \mathrm{Dom}(F_l) \setminus a_l^{(r)}$ such that $a_l^{(r)} = x_{il}$

$$D_{ik}^{GH} = \sum_{l=1}^{p} \delta(x_{ik}, v_{kl}) \tag{4.4}$$

where

$$\delta(x_{ik}, v_{kl}) = \sum_{t \in \mathcal{D}_{il}} w_{kl}^{(t)} = \sum_{t \in \mathcal{D}_{il}} \frac{1}{|A_k|} \sum_{\omega_\nu \in A_k} w_{\nu l}^{(t)}. \tag{4.5}$$

The factor $\frac{1}{|A_k|}$ in (4.5) is used to penalize subsets of clusters with high cardinality.

Remark 12. A normalized version of the new generalized Hamming distance, i.e., $D_{ik}^{GH} \in [0, 1]$ can be considered by multiplying D_{ik}^{GH} by $\frac{1}{p}$.

In the next section, the objective function, updating formula of the centers, and evidential partition are presented.

4.3 cat-ECM: categorical evidential c-means

In this section, we present the objective function of *cat-ECM*, the updating formula of the evidential partition and the cluster centers, and the algorithm of *cat-ECM*.

Objective function

The objective function of *cat-ECM* remains the same as *ECM* in Equation (2.16) by replacing the squared distance with the new generalized Hamming distance and considering the centers defined in Section 4.2.

Optimization

Following the alternate optimization scheme, when the variable $\mathbf{V}$ is fixed the updating formula of the evidential partition is given by Equation (2.18) and (2.19). When $\mathbf{M}$ is fixed, the updating formula of the attributes category weights is given by the following equation:

$$w_{\nu l}^{(t)} = \begin{cases} 1 \text{ if } & f_{i\nu}^{(t)} = f_{i\nu}^{(r)} \text{ with } r = \underset{z\in\{1,\dots,n_l\}}{\arg\max}\, f_{i\nu}^{(z)} \\ \frac{1}{q} \text{ if } & \exists s_1,\dots,s_{q-1} s.t. f_{i\nu}^{(t)} = f_{i\nu}^{(s_1)} = \cdots = f_{i\nu}^{(s_{q-1})} > f_{i\nu}^{(z)}, \\ & \forall z \in \{1\dots n_l\},\ s.t.\ z \neq s_1 \neq \cdots \neq s_{q-1} \neq t, \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

$\forall \nu \in \{1,\dots,c\},\ \forall l \in \{1,\dots,p\},\ \forall t \in \{1,\dots,n_l\}$, with

$$f_{i\nu}^{(t)} = \sum_{A_k\subseteq\Omega, \omega_\nu\in\mathbf{A}_k} \sum_{i/\mathbf{x_{il}}=a_l^{(t)}} |\mathbf{A}_k|^{\alpha-1} m_{ik}^{\beta}.$$

Proof. Let $D_{il} = \mathrm{Dom}(F_l) \setminus a_l^{(r)}$ and $a_l^{(r)} = \mathbf{x_{il}}$. We have

$$\begin{aligned} J_{cat-ECM}(\mathbf{M},\mathbf{V}) &= \sum_{i=1}^{n} \sum_{\mathbf{A}_k\subseteq\Omega} |\mathbf{A}_k|^{\alpha} m_{ik}^{\beta} D_{ik}^{GH} + \sum_{i=1}^{n} \rho^2 m_{i\emptyset}^{\beta} \\ &= \sum_{i=1}^{n} \sum_{\mathbf{A}_k\subseteq\Omega} |\mathbf{A}_k|^{\alpha-1} m_{ik}^{\beta} \sum_{l=1}^{p} \sum_{t\ \in D_{il}} \sum_{\omega_\nu\in\mathbf{A}_k} w_{\nu l}^{(t)} + \sum_{i=1}^{n} \rho^2 m_{i\emptyset}^{\beta}. \end{aligned}$$

Since

$$\sum_{\mathbf{A}_k\subseteq\Omega} \sum_{\omega_\nu\in\mathbf{A}_k} |\mathbf{A}_k|^{\alpha-1} m_{ik}^{\beta} = \sum_{\nu=1}^{c} \sum_{\omega_\nu\in\mathbf{A}_k, A_k\subseteq\Omega} |\mathbf{A}_k|^{\alpha-1} m_{ik}^{\beta}, \tag{4.7}$$

the objective function can be written as:

$$J_{cat-ECM}(\mathbf{M},\mathbf{V}) = \sum_{\nu=1}^{c} \sum_{l=1}^{p} \sum_{i=1}^{n} \sum_{\omega_\nu\in\mathbf{A}_k, A_k\subseteq\Omega} |\mathbf{A}_k|^{\alpha-1} m_{ik}^{\beta} \sum_{t\ \in D_{il}} w_{\nu l}^{(t)} + \sum_{i=1}^{n} \rho^2 m_{i\emptyset}^{\beta}.$$

Let $w_{\nu l}^{(r)}$ be the weight associated to the attribute value equal to $\mathbf{x_{il}}$. Using (4.2), we deduce that $\sum_{t\ \in D_{il}} w_{\nu l}^{(t)} = 1 - w_{\nu l}^{(r)}$. Thus,

$$J_{cat-ECM}(\mathbf{M},\mathbf{V}) = \sum_{\nu=1}^{c} \sum_{l=1}^{p} \sum_{i=1}^{n} \sum_{\omega_\nu\in\mathbf{A}_k, A_k\subseteq\Omega} |\mathbf{A}_k|^{\alpha-1} m_{ik}^{\beta} \left(1 - w_{\nu l}^{(r)}\right) + \sum_{i=1}^{n} \rho^2 m_{i\emptyset}^{\beta}.$$

As $\mathbf{M}$ is fixed, the terms $\sum_{i=1}^{n} \rho^2 m_{i\emptyset}^{\beta}$ and $\sum_{i=1}^{n} \sum_{\omega_\nu \in \mathbf{A}_k, A_k \subseteq \Omega} |\mathbf{A}_k|^{\alpha-1} m_{ik}^{\beta}$ are constant. Since each element $w_{\nu l}$ are independent, minimizing $J_{cat-ECM}(\mathbf{V})$ is equivalent to maximizing (4.8) under the same conditions.

$$J_1(w_{\nu l}) = \sum_{i=1}^{n} \sum_{\omega_\nu \in \mathbf{A}_k, A_k \subseteq \Omega} |\mathbf{A}_k|^{\alpha-1} m_{ik}^{\beta} w_{\nu l}^{(r)}, \quad \forall \nu \in \{1, \ldots, c\},\ \forall l \in \{1, \ldots, p\} \tag{4.8}$$

Taking for all objects and all subsets w_ν the values $|\mathbf{A}_k|^{\alpha-1} m_{ik}^{\beta}$ and the weight associated to $\mathbf{x_{il}}$ is similar to taking separately each possible weight $w_{\nu l}^{(r)}$ of the attribute F_l and summing the values $|\mathbf{A}_k|^{\alpha-1} m_{ik}^{\beta}$ associated to objects having the same value $a_l^{(r)}$ and subsets containing ω_ν. This leads to write $J_1(w_{\nu l})$ as follows:

$$J_1(w_{\nu l}) = \sum_{t=1}^{n_l} w_{\nu l}^{(t)} \underbrace{\sum_{\omega_\nu \in \mathbf{A}_k, A_k \subseteq \Omega} \sum_{i/\mathbf{x_{il}} = a_l^{(t)}} |\mathbf{A}_k|^{\alpha-1} m_{ik}^{\beta}}_{\text{constant}}, \quad \forall \nu \in \{1, \ldots, c\},\ \forall l \in \{1, \ldots, p\} \tag{4.9}$$

The two last sums correspond to a constant, thus, the maximization of the objective function under the constraints (4.1) and (4.2) is a linear optimization problem with linear constraints similarly to the problem (3.5). An optimal solution to this problem is given by Equation (4.6). □

Remark 13. In the left term of (4.7), the sum over all ω_k are considered at the same time while in the right term, they are considered independently.

It can be noted that in practice most of the time the obtained weights of categories in single clusters are crisp. Consequently the weights of categories in subsets of clusters $\mathbf{A}_k$ with cardinality greater than one, i.e., $|A_k| > 1$ are $\frac{1}{|\mathbf{A}_k|}$.

Algorithm and complexity analysis

An algorithm for the proposed method can be derived as follows: from the input parameters, the centers of singletons are initialized such that Equation (4.1) and (4.2) are satisfied. Then the centers of subsets of clusters with cardinality > 1 are computed by taking the means of category weights of singletons. In the next step, the evidential partition and the centers are

updated respectively with Equations (2.18) and (2.19) and (4.6). The algorithm is summarized in Algorithm 9.

To compute $f_{i\nu}^{(t)}$, in the worst case, $O(cn)$ operations are needed. For c single clusters, the time complexity is $O(c^2 n\mathcal{J})$. For subsets of clusters such that $|A_k| > 1$, the average of single clusters weights is computed. In the worst case, i.e., $A_k = \Omega$, c averages are computed for $\mathcal{J}$ categories. As there are $2^{c-1} - c$ subsets with $|A_k| > 1$, the time complexity for these subsets is then $O(c2^{c-1}\mathcal{J}) \rightarrow O(c2^c\mathcal{J})$. Hence, the overall time complexity to update the weights is $O(c^2 n\mathcal{J} + c\varrho\mathcal{J})$ when $\varrho = 2^c$, i.e., in the worst case. For the distance, the time complexity $O(n\varrho\mathcal{J})$. For the partition matrix, it is the same as *ECM*'s, i.e., $O(n\varrho)$. Consequently the overall time complexity is $O(c^2 n\mathcal{J} + c\varrho\mathcal{J} + n\varrho\mathcal{J} + n\varrho) \rightarrow O(c^2 n\mathcal{J} + c\varrho\mathcal{J} + n\varrho\mathcal{J})$ for one iteration. By considering T iterations, the time complexity is $O(c^2 n\mathcal{J}T + c\varrho\mathcal{J}T + n\varrho\mathcal{J}T)$.

The memory complexity of *cat-ECM* is given by $O(np + n\varrho + \varrho\mathcal{J})$ as $\mathbf{X}$ is $O(np)$, D is $O(n\varrho)$, $\mathbf{U}$ is $O(n\varrho)$ and $\mathbf{V}$ is $O(\varrho\mathcal{J})$.

Algorithm 9 Categorical evidential c-means algorithm (cat-ECM)

Require: $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_i}, \ldots, \mathbf{x_n}\}$ the categorical data, $1 < c < n$ the number of clusters, $\alpha \geq 1$ the weighting exponent for cardinality, $\beta > 1$ weighting exponent, $\rho > 0$ the distance to the outliers cluster, and $\epsilon \geq 0$ a threshold.
Output: $\mathbf{M}$ evidential partition of $\mathbf{X}$ and $\mathbf{V}$ the centers of clusters.
Begin
  Randomly initialize $\mathbf{V}_0$ that respects (4.1) and (4.2)
  $\tau \leftarrow 0$
  repeat
    $\tau \leftarrow \tau + 1$
    Compute the distance matrix with the new generalized Hamming distance in (4.4).
    Update M_τ using (2.18) and (2.19) with $d_{ik}^2 = D_{ik}^{GH}$.
    Update clusters center V_τ using (4.6).
  until $\|\mathbf{J}_{cat-ECM}^{\tau} - \mathbf{J}_{cat-ECM}^{\tau-1}\| \leq \varepsilon$
End

In the next section, we describe the protocol of the experiments.

4.4 Experiments

This section presents the datasets used in experiments, the evaluation criteria used to test the performance of *cat-ECM* and compares the latter to existing methods, the experimental protocol, the materials used to conduct experiments. Finally, the results of the experiments are presented.

4.4.1 Data sets

We use the same datasets (Soybean, Zoo, Breast, Lung, Votes, Credits, Cars, Dermatology, Mushrooms) presented in Section 3.4.1 to conduct experiments. In addition, we use the categorical version of the Butterfly dataset in Figure 4.1 to illustrate the strengths of *cat-ECM*.

4.4.2 Evaluation criteria

Internal measures

Like in experiments of *CFE* in Chapter 3, we use the PC, PE, and FS as internal measures. Due to the evidential partitions obtained from *cat-ECM* and *ECM* we also use the nonspecificity (N) in Equation (1.25) to quantify and compare the uncertainty in these partitions.

External measures

In addition to the RI and FRI, we use the credal rand index (CRI) introduced in [89]. The CRI is an extension of the RI and FRI in the belief functions framework. It allows the comparison of different types of partitions such as hard, fuzzy, possibilistic [90] and evidential.

Let $\mathbf{M}$ and $\mathbf{M}'$ be two partitions to be compared and $\mathcal{R} = (m_{ij})_{1\leq i\leq j\leq n}$ and $\mathcal{R}\prime = (m\prime_{ij})_{1\leq i\leq j\leq n}$ be respectively their relational representations [89]. The CRI is defined as follows:

$$CRI = 1 - \frac{2\sum_{i<j}\varphi(m_{ij}, m\prime_{ij})}{n(n-1)}, \tag{4.10}$$

where φ is a distance between mass functions and n the number of objects. In [89], the authors proposed two versions of the CRI where φ corresponds to Jousselme's [91] (CRI^J) and belief [92] (CRI^B) distances.

The CRI varies from 0 to 1 with 1 corresponding to the optimal value. When the true classes of objects are known, the consistency of the evidential partitions with the true classes can be computed by considering the degree of conflict (1.20) as φ in Equation (4.10). Therefore, in this case, the consistency is more useful than the CRI^J and CRI^B.

Based on the credal index CRI and the nonspecificity N, an evidential partition $\mathbf{M}$ is said preferable to another one $\mathbf{M}'$ when $\mathbf{M}$ is more consistent with the true classes and more precise [89], in other words, when

$$CRI(\mathbf{M}) > CRI(\mathbf{M}') \text{ and } N(\mathbf{M}) < N(\mathbf{M}'). \tag{4.11}$$

We use the two measures to compare evidential partitions in experiments.

Partitions	Descriptions
TRUE	Vector of real classes
k-modes	Hard partition from *k-modes*
FC*-H	Hard partition from *FC**
FC*-F	Fuzzy partition from *FC**
CFE-H	Hard partition from *CFE*
CFE-F	Fuzzy partition from *CFE*
cat-ECM-H	Hard partition from *cat-ECM*
cat-ECM-F	Fuzzy partition from *cat-ECM*
cat-ECM-C	Evidential partition from *cat-ECM*
ECM-H	Hard partition from *ECM*
ECM-F	Fuzzy partition from *ECM*
ECM-C	Evidential partition from *ECM*

Table 4.1: Compared partitions based on the consistency with the true classes.

4.4.3 Experimental protocol

We run experiments on the nine datasets following the protocol described in Figure 3.2. We compare the performance of *cat-ECM* against *ECM*, *CFE*, *FCM*, *FC**, *k-modes* following the same procedures as in Chapter 3. In all experiments, we transform the categorical datasets to numerical with the one-hot encoding technique in order to use the numerical methods *ECM* and *FCM*.

In addition to the performance comparisons, we also compare the partitions obtained from the preceding methods using the CRI and nonspecificity as proposed in [89]. As the evaluation criteria PE, PC, RI, FRI, and FS take into account hard and fuzzy partitions, we generate these partitions from evidential methods and hard partitions from fuzzy methods as follows:

- For evidential models (*cat-ECM* and *ECM*) we normalize the evidential partitions with Dempster normalization function in Equation (1.13) and compute the fuzzy partitions with the pignistic transformation in Equation (1.23). From the latter partitions, we generate hard partitions with the maximum principle rule i.e. the objects are assigned to the cluster with the highest fuzzy membership degrees.
- For fuzzy models (*FC** and *FCM*), the hard partitions are obtained as preceding with the maximum principle rule.

The obtained partitions are summarized in Table 4.1. In the next subsections, present the parameter settings and materials used in experiments.

4.4.4 Parameter settings

In all experiments, we set the values of the parameter β in *cat-ECM* and *ECM* which corresponds to m in *CFE* from 1.1 to 2. In addition, we keep the default value of the parameter ρ in the implementation of *ECM* and set the same value for *cat-ECM*. To reduce the time complexity of the methods, we consider only subsets corresponding to the empty set, the subsets of size less than 2 and Ω.

4.4.5 Materials

We use the same materials as in Chapter 3 to conduct experiments. For the partitions comparisons, we use the implementations of *ECM* and *CRI* provided by Antoine Violaine[1]
We provide the code source for the implementation of *cat-ECM* on Github[2].
In the next section, the results of the experiments are presented.

4.4.6 Results

This section describes the results of the experiments from the protocol in Section 4.4.3. We firstly discuss the scores of RI, FRI, FS, PC, and PE obtained from *cat-ECM*. Secondly, we present the results of the statistical comparisons of the models and finally we present the partitions comparisons results.

cat-ECM results

Scores of RI, FRI, FS, PC, PE obtained with *cat-ECM* through the protocol 4.4.3 are summarized respectively in Tables B.1.1, B.1.2, B.1.3, B.1.4 and B.1.5 from Appendix B. We can first note that contrary to *CFE*, there is no continuous decrease in the scores when the value of β increases. From the preceding tables, it can be seen that optimal scores over the datasets are obtained with different values of β. For the FRI, PE, and PC scores, optimal values are obtained for $\beta = 1.1$. For the FS and RI scores the optimal values of β are not unique. Similarly to *CFE*, it can be noted that the scores PC and PE behave similarly as explained in Chapter 3.

Statistical comparisons results

From the statistical comparisons of the models over the scores and the datasets, we obtain the following results. For all the values of $\gamma \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7\}$ the difference of the performances of the models is not statistically significant for almost the RI, FRI and FS scores.

[1] violaine.antoine@uca.fr
[2] https://github.com/abdjiber/catecm

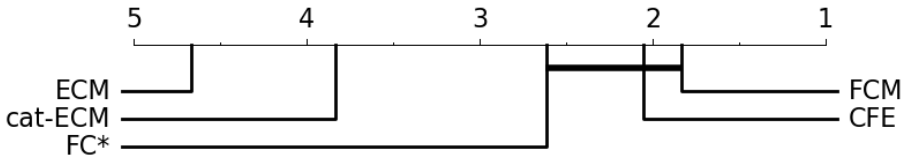

Figure 4.2: Critical difference diagram obtained for PC and PE scores with $\gamma = 0.1$ and $\beta = 1.1$.

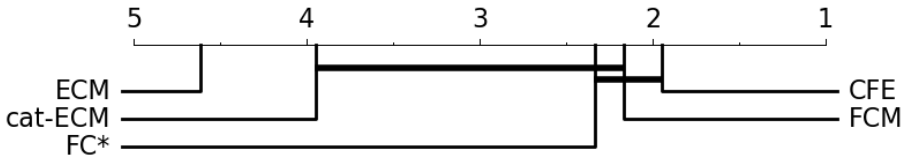

Figure 4.3: Critical difference diagram obtained PC scores with $\gamma = 0.1$ and $\beta = 1.2$.

For the PC and PE scores, a statistical significance is obtained for $\gamma = 0.1$ and $\beta \in \{1.1, 1.2\}$. The Critical difference diagram for $\gamma = 0.1$ and $\beta = 1.1$ of these scores is represented in Figure 4.2. In this figure, the ranks of the models in order are respectively *FCM*, *CFE*, *FC**, *cat-ECM* and *ECM*. While the pairwise comparisons between *CFE*, *FCM* and *FC** is not statistically significant, the test is significant between the other models. it can be noted that based on these scores, the FCM-like models have the highest performance and the ECM-based models the lowest.

Figures 4.3 and 4.4 correspond respectively to the critical difference diagrams of PC and PE when $\gamma = 0.1$ and $\beta = 1.2$. The difference between the two figures holds on the ranking of *FC** and *FCM*. While the ranks are different in the former figure, they are the same in the latter one. For the FRI scores, the Friedman test is significant for $\gamma = 0.5$ and $\beta = 1.1$, the corresponding Critical difference diagram is represented in Figure 4.5. In the latter figure, all the models statistically outperformed *ECM*. As the tests are not significant for the RI and FS scores for all the considered values of γ, we provide in Appendix B.2 the critical difference diagrams for the highest and lowest value of β (1.1, and 2) and $\gamma = 0.1$. In the next subsection, we discuss the results of the partitions comparisons based on the consistency with the true classes.

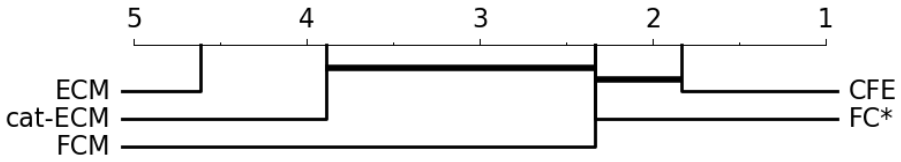

Figure 4.4: Critical difference diagram obtained PE scores with $\gamma = 0.1$ and $\beta = 1.2$.

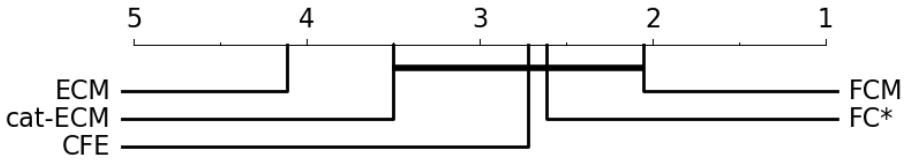

Figure 4.5: Critical difference diagram obtained FRI scores with $\gamma = 0.5$ and $\beta = 1.1$.

Partitions comparisons results

For the comparisons of the partitions described in Table 4.1, we select the datasets in Table 3.2 with the smallest and highest number of categories (Zoo, Votes, and Credits). Since the *CFE*, *cat-ECM*, *FC** and *k-modes* algorithms handle categorical data without encoding as for *ECM* and *FCM*, the idea of the selection of these datasets is to compare the consistency of the partitions with the true classes. When the number of categories is small, i.e., with the one-hot encoding technique, the original categorical dataset is slightly changed by the addition of few dimensions whereas for a higher number of categories, the dimension of the original dataset will be drastically changed. We also consider in the comparisons the cases when the value of the parameter β are the smallest (1.1) and the highest (2).

Remark 14. It should be noted that the nonspecificity of all non-evidential partitions is equal to 0.

Figures 4.6, 4.7 and, 4.8 correspond respectively to the results of comparisons of the partitions on the Zoo, Votes and Credits dataset when $m = 1.1$. We describe below these results by the types of partitions (Evidential - Evidential, Evidential - Fuzzy - Hard, Fuzzy - Hard) as follows:

- **Evidential - Evidential**: In Figures 4.6, 4.7, and 4.8, it can be noted that based on Equation (4.11), the evidential partition of *cat-ECM* is preferable to the one of *ECM*. In Section B.3 of Appendix B, we provide the results of comparisons on the Zoo, Votes and Credits

datasets of the partitions when $m = 2$. With this configuration, the evidential partition of *ECM* is preferable to *cat-ECM*.

- **Fuzzy - Hard - Evidential**: In all the three preceding figures, the consistencies of *cat-ECM* and *ECM* are higher than those of all the fuzzy and hard partitions.

- **Hard - Fuzzy**: In Figure 4.6, it can be noted that the consistencies of the fuzzy and hard partitions of *CFE* (not visible in the Figure: 0.93 for both partitions) are lower than those of *FC**'s partitions (0.94 for *FC*-H* and 0.95 for *FC*-F*) and *k-modes*'s (0.94) but higher than those obtained with the partitions of *FCM* (0.92 for both *FCM-H* and *FCM-F*). In Figure 4.7, while the consistency of *ECM-F* is the lowest, the one of *ECM-H* is the same as *FCM-H* and greater than those of *FC**'s, *CFE*'s and *k-modes*'s. In Figure 4.8, the consistencies of *FCM-F*, *FCM-H*, and *ECM-F* are about 0.5 whereas the ones of *CFE-F*, *CFE-H*, *FC*H* and *k-modes* are about 0.67.

  From the consistencies in Section B.3 from Appendix B, the *k-modes* partition obtain the third-highest consistency and the hard partition of *CFE* the lowest on the Zoo dataset. On the Votes datasets, the consistencies of the fuzzy partitions of *ECM* and *FCM* are the lowest while their hard partitions have the third-highest consistencies (around 0.79 for both partitions). For the Credits dataset, the consistency of the hard partition of *CFE* is higher than all of the other partitions except the one of *k-modes*.

In the next section, we illustrate through the ordinal categorical Butterfly and the Zoo datasets, the strengths of *cat-ECM*.

4.5 Strengths of cat-ECM

By using the Dempster Shafer theory and a noise clustering objective function *cat-ECM* allows capturing overlapping objects (the ones belonging to subsets of clusters) and outliers. In this section, we conduct some experiments on the Butterfly and Zoo dataset to illustrate these behaviors.

In the first experience, we run *cat-ECM* on the categorical Butterfly dataset in Figure 4.1 with the following parameters $c = 2$ ($\Omega = \{C_1, C_2\}$), $\beta = 1.1$, $\alpha = 1$, and $\delta = 1.2$. We ran the algorithm 30 times and selected the evidential partition at the minimum of the cost function. For each object we plot the mass corresponding to the highest cluster membership degree and obtained Figure 4.9.

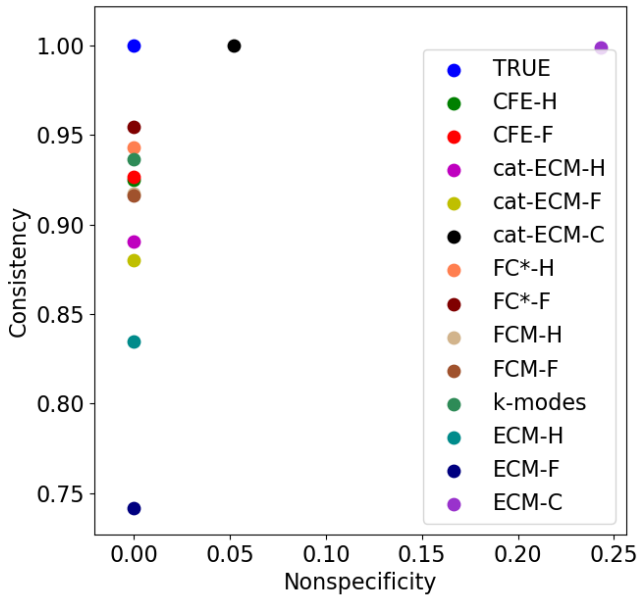

Figure 4.6: Nonspecificity against Consistency obtained on the Zoo dataset with $\beta = 1.1$.

From Figure 4.9, it can be noted that the set of objects $\{o2, o3, o4\}$, $\{o8, o9, o10\}$, $\{o1, o5, o6, o7, o11\}$ and $\{o12, o13\}$ are respectively assigned to cluster C_2, C_1, Ω and the empty set. In Section 4.1, we expected objects $\{o1, o2, o3, o4\}$ and $\{o8, o9, o10, o11\}$ to be in the same clusters, objects $\{o5, o6, o7\}$ to be in Ω and objects $\{o12, o13\}$ to be outliers ($\in \emptyset$).

Apart $o1$ and $o11$, *cat-ECM* successfully assigned the objects into the expected clusters. By ignoring the order, objects $\{o1, o5, o6, o7, o11\}$ have similar characteristics such as the same coordinate y. Consequently, based on this similarity, these objects are expected to belong to the same cluster, in our case Ω. It should be noted that the misassignments of objects o_1 and $o11$ can be due to inappropriate parameter settings of *cat-ECM*.Finally, objects $o2$ and $o8$ are assigned in C_1 and C_2 with full certainty, therefore these objects correspond to the clusters of the centers.

In the second experiment, we select the Zoo dataset to illustrate again the ability of *cat-ECM* to capture subsets of clusters. The selection of this dataset among the nine is due to the simplicity of interpretation of the obtained clusters. We use the results of the principal components analysis in Figure 3.1 and illustrative images of each animal of the dataset to plot

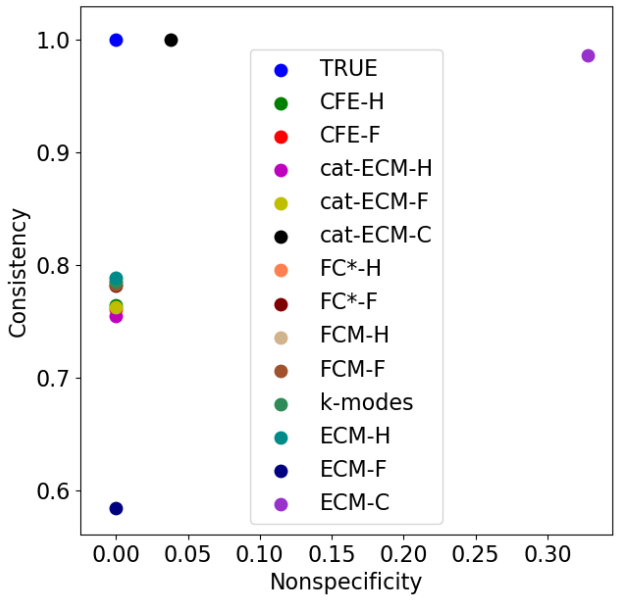

Figure 4.7: Nonspecificity against Consistency obtained on the Votes dataset with $\beta = 1.1$.

Figure 4.10. We run the *cat-ECM* algorithm with the following parameters: $\alpha = 1.5$, $\beta = 1.1$, $\delta = 10$ over ten iterations and took the evidential partition corresponding to the lowest cost function. From the evidential partitions, we assign the objects to the cluster with the highest membership degree. In Figure 4.10, we plot the obtained clusters and to each object's coordinates, we associate an image of the corresponding animal.

In the latter figure, cluster C1 mostly represents amphibians and reptiles, C2 aquatic invertebrates, C3 aquatic mammals, C4 insects, C5 land mammals, C6 birds, and C7 fishes. The subset C56={land mammals, birds} corresponds to the fruit bat and vampire. These two animals are of the family of bats and constitute the only flying mammals in the dataset. Therefore their assignment to this subset makes sense.

The subset C35={aquatic mammals, land mammals} corresponds to the mink which is a semi-aquatic mammal. The subset C16={amphibian, birds} corresponds to the penguin. As it is an aquatic flightless bird it can be assigned with certainty neither in C1 nor in C6. The subset C17={reptiles, fishes} corresponds to the pitviper and sea snake which is are reptiles and can be aquatic animals, they then share similar characteristics as swimming with

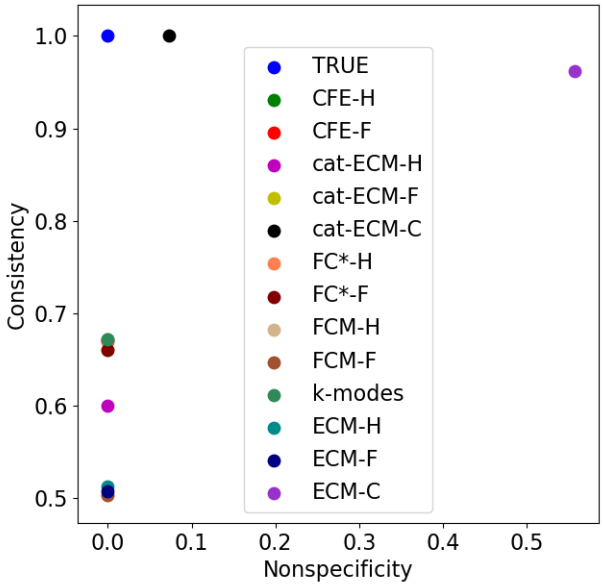

Figure 4.8: Nonspecificity against Consistency obtained on the Credits dataset with $\beta = 1.1$.

fishes. In the dataset, the attribute *aquatic* is set 1 and 0 respectively for the sea snake and pitviper. Hence, there is an uncertainty to fully assign them to C1 or C7. The subset C13={amphibian, aquatic mammals} corresponds to the platypus which is a semi-aquatic mammal.

The subset C15={reptiles, land mammals} corresponds to the tortoise which is a reptile. Based on the features in the dataset, the assignment of the tortoise in C5 can be explained by the fact that it has a tail like all the land mammals of the dataset. Finally, the subset $\mathbf{\Omega}$ corresponds to the scorpion. The assignment of this animal to $\mathbf{\Omega}$ represents a complete ignorance about its subset. This observation can be explained by the fact that the scorpion shares similar characteristics with most of the animals in the dataset. Indeed the scorpion is a predator, breathes, and has a tail as respectively 55%, 79% and 75% of the animals in the dataset.

In the upcoming section, we describe some limitations of the *cat-ECM* algorithm.

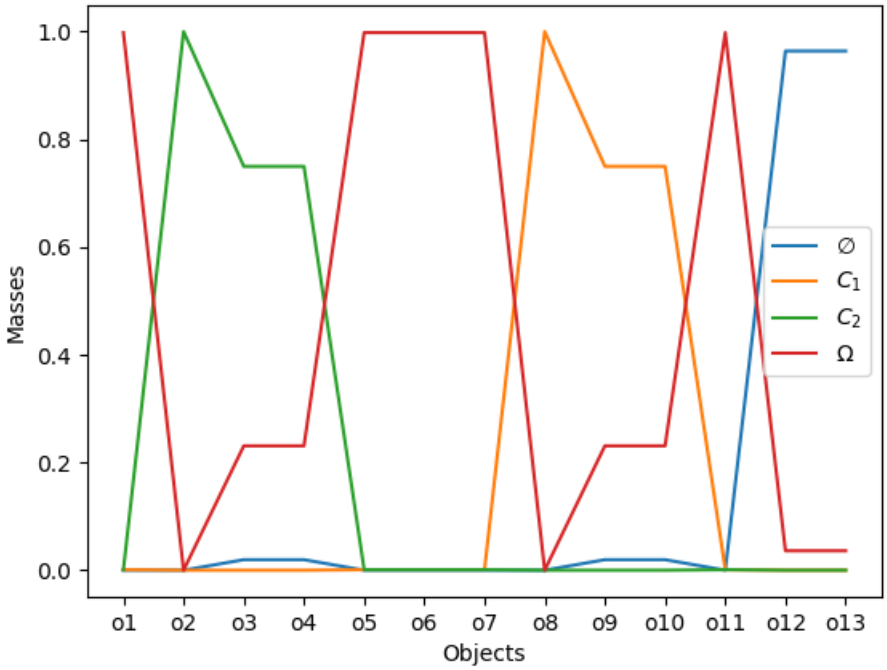

Figure 4.9: Mass obtained with *cat-ECM* on the ordinal categorical Butterfly dataset.

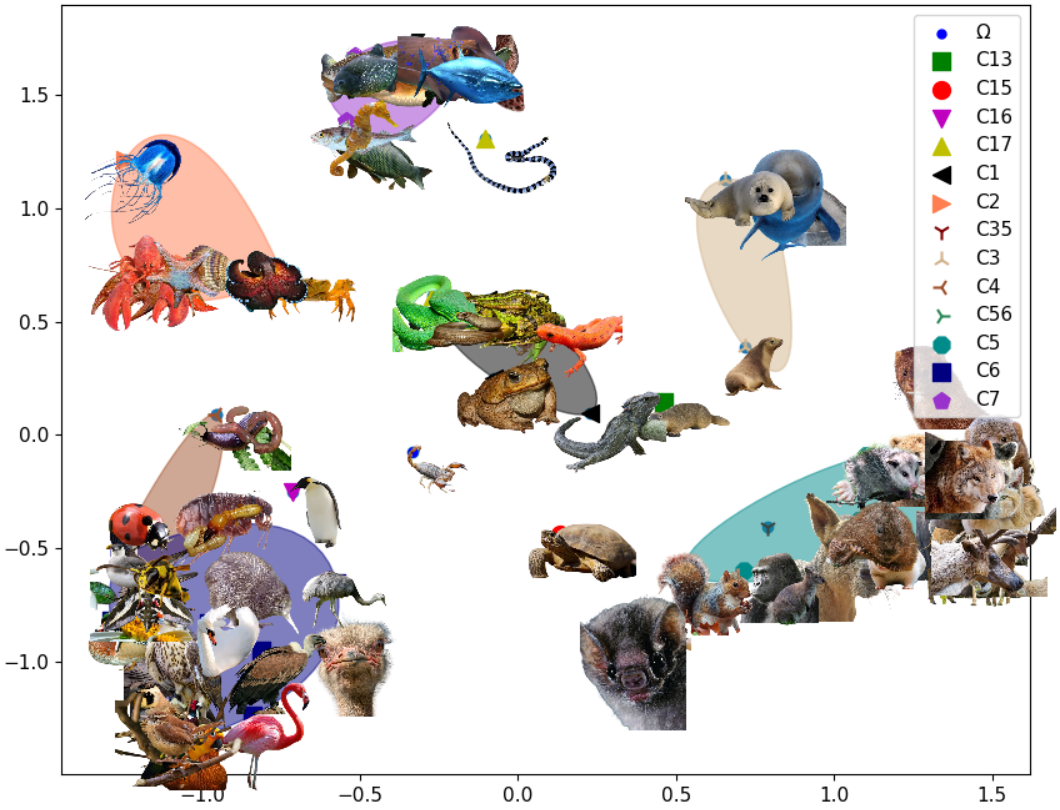

Figure 4.10: Subsets of clusters obtained with *cat-ECM* on the Zoo dataset.

4.6 Limitations of cat-ECM

In this section, we present some limits of *cat-ECM* namely, the setting of optimal values of β and α and the time complexity of the method.

4.6.1 Optimal values of β and α

As an extension of FCM-like algorithms, *cat-ECM* inherits most of the limits of FKM, particularly the setting of optimal value of β (m for *CFE*) discussed in Section 3.6. In addition to this parameter, the setting of an optimal value of α and δ is also problematic. As noted in [59], for the value of δ, Davé suggested in [93] an iterative value based on the average distance between the objects and the clusters as follows

$$\delta_\tau = \frac{\rho}{nc} \sum_{i=1}^{n} \sum_{j=1}^{c} d_{ik}^2, \tag{4.12}$$

where τ is the corresponding iteration and ρ a user input parameter corresponding to the multiple of the average distance
Davé's solution for the setting of an optimal value of δ is used neither in our theoretical work nor in the experiments.

As $|A_k|^\alpha$ is a penalization term of subsets with high cardinality, a higher (resp. lower) value of α leads to a high (resp. low) penalization. Similarly, regarding the fuzzy entropy coefficient of *CFE*, there is no formal way to determine an optimal value of α. Despite the setting of this value to 2 in most of our experiments for simplicity, we recommend conducting experiments to determine an optimal value.

Due to the use of subsets of clusters, *cat-ECM* may require high computational resources. We discuss this limitation in the next subsection.

4.6.2 Time complexity

The time complexity analysis in Section 4.3 shows that the time complexity of *cat-ECM* is $O(c^2 n\mathcal{J}T + c\varrho\mathcal{J}T + n\varrho\mathcal{J}T)$. Depending on the value of ϱ, the latter complexity can be an exponential function of the number of clusters, i.e., when $\varrho = 2^c$. For instance, for the Zoo dataset, $2^7 = 128$ subsets will be used. To overcome this issue, a solution would be to limit the number of subsets. A common choice is to set the subsets to the empty set, singletons, pairs, and Ω. By default, the implementation of *cat-ECM* in the open-source takes into account this choice.

Conclusion

We introduce in this chapter a categorical extension of the evidential c-means algorithm. The new method referred to as *cat-ECM* uses a generalized Hamming distance and fuzzy centers. We follow the experimental protocol discussed in chapter 3 to evaluate the performance of the model. To that end, we use different values of the input parameters of *cat-ECM* and compare its performance to *CFE*, *FC**, *k-modes*, *FCM* and *ECM* with internal and external measures namely RI, FRI, FS, PE and PC. From the obtained results, in most of the cases, the statistical difference between the models is not significant. For the parameter settings where the statistical tests were significant, *CFE* usually has the highest rank on the critical diagrams. In the next step, we compare the consistency of the hard, fuzzy, and evidential partitions generated from the models with the true labels on three selected datasets with the lowest and highest numbers of attributes categories. For the evidential partitions obtained with *cat-ECM* and *ECM*, in addition to the consistency we compare their nonspecificity. The results of these comparisons suggest that *cat-ECM* works better with small values of β whereas, for *ECM*, good performances are obtained in general with high values of β.

Contrary to FCM-like clustering models, *cat-ECM* provides through the Dempster Shafer theory more flexible ways to capture uncertainty on object assignments to clusters. We illustrate this behavior on the categorical Butterfly and Zoo datasets. Moreover, *cat-ECM* achieve the performance of numerical methods such as *ECM* and *FCM* and sometimes better by handling categorical data without transformations.

As with many clustering methods, *cat-ECM* has some limitations already known in the literature for not having a formal way of solving them. For example, the setting of optimal values of β and α. Nonetheless, for the parameter β, Davé proposed in [93] an iterative procedure to determine a convenient value. In addition to the preceding limitations, due to the use of subsets of clusters, the time complexity of *cat-ECM* can grow exponentially. To overcome this problem, the subsets of clusters generated from *cat-ECM* can be limited to for instance the empty set, the singletons, pairs, and Ω.

In the next chapter, we present real-world applications of *CFE* and *cat-ECM* to study the influence of lifestyle activities on cognitive health.

Key points

- *cat-ECM* is an extension of the *ECM* algorithm for categorical data. The method generates subsets of clusters, allowing to capture uncertainties from overlapped objects and outliers.
- In the experiments, in most of the cases, the statistical analysis that compared the performances of *cat-ECM* and existing methods were not significant.
- We illustrate the strengths of *cat-ECM* on the Butterfly and Zoo datasets and discussed some of the limitations of the method.
- The time and memory complexity of *cat-ECM* are respectively $O(c^2 n\mathcal{J}T + c\varrho\mathcal{J}T + n\varrho\mathcal{J}T)$ and $O(np + n\varrho + \varrho\mathcal{J})$.

Publications and communications

The *cat-ECM* algorithm was presented at the IEEE International Conference on Fuzzy Systems in 2019. We provide below a reference to the paper.

- A. J. Djiberou Mahamadou, V. Antoine, G. J. Christie, and S. Moreno, "Evidential clustering for categorical data," 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, LA, USA, 2019.

Part III

Applications

5 — Differential susceptibility in older adults: cluster analysis perspectives

Contents

5.1 Background 122
5.1.1 Lifestyle factors and cognitive health in older adults . 122
5.1.2 Orchid and dandelion theory 124
5.1.3 Soft computing in medicine and healthcare 127
5.2 Experiments 127
5.2.1 Data set 128
5.2.2 Evaluation criteria 129
5.2.3 Experimental protocol 132
5.2.4 Parameter settings 132
5.2.5 Materials 132
5.2.6 Results 132
5.3 Discussions 142
5.4 Conclusion 142

Introduction

Every 3 second, someone in the world develops Dementia[1] [94]. Consequently, an exponential increase of the total cost of the diseases related to

[1] Dementia is a collection of neurodegenerative diseases. Alzheimer's disease is among the most frequent.

Dementia is observed: from 818 billion dollars in 2015 to a trillion in 2018 [94]. With the rapid growth of the older adults population worldwide (1 in 6 people in the world will be over the age of 65 by 2050, up from 1 in 11 in 2019 according to the United Nations [95]) from which 152 million are expected to develop Dementia by 2050 [96], public health policymakers and healthcare systems face many challenges. It then becomes an urge to better understand the mechanism of older adults aging and prevent their cognitive decline. To this end, researchers have conducted several studies in which different research questions on aging were explored. Among the studies, the impact of environment (lifestyle activities) on cognitive health in older adults received much interest. Indeed, several studies have shown in the literature the critical role played by the environment in the preservation or the decline of cognitive health. In this chapter, we conduct a complementary analysis based on the developed clustering methods to investigate the replication of recent findings on the influence of environmental conditions on older adults.

In the next section, we provide a background on the literature on the influence of environmental conditions on cognitive health, a review of the differential susceptibility theory, and some applications of soft computing in medicine and health.

5.1 Background

Over decades, cognitive decline has become a growing public health issue. The risk factors related to cognitive decline can be summarized into two groups: non-modifiable and modifiable factors. The first group of risk factors characterized by age and genes constitutes the factors that cannot be modified by medical interventions or individual behavior. Among the modifiable risk factors, several lifestyle factors, i.e. environment, have been identified such as education, smoking, physical activities, diet, alcohol, cognitive training, depression, and sleep (see [97] for example for a review). In [98], Christie et al. highlight the importance of basic lifestyle activities such as physical exercise, mediation, and musical experience to slow down the progression of Dementia in older adults as there are low cost, easily scalable, and can be brought to market quickly.

In the next subsection, we review some papers that studied the influence of lifestyle factors on cognitive health.

5.1.1 Lifestyle factors and cognitive health in older adults

To assess the interactions between environmental factors and cognitive health, in data-driven studies (i.e., involving statistical models), memory measures

are usually used as dependent (outputs) variables and environmental factors as independent variables. Indeed, memory function has been shown to decline with age [99]. In the literature, a panel of memory assessment tests has been proposed (see [100] for a review). We review below some studies on the impact of lifestyle activities on cognitive health.

In [101], the authors used the longitudinal clinical pathologic data from the Rush Memory and Aging Project [102] which contains 1148 individuals with a mean age and education of respectively 80.4 and 14.5 years, to test the hypothesis that late-life participation in mentally stimulating activities affects subsequent cognitive health. The assessment of cognitive function in the study was carried out with 20 individual performance tests, including the Mini-Mental State Examination (MMSE) [103]. A composite measure of global cognition is created with 19 tests. The statistical analysis results with cross-lagged panel models [104, 105] suggest that more frequent mental stimulation in old age leads to better cognitive functioning. In [106], the authors investigated the role of cognitive reserve [2] and lifestyle factors early in life on healthy aging. More specifically, the authors studied three questions about the cognitive reserve: (1) Does cognitive reserve follow a static or dynamic change pattern across the lifespan? 2) can cognitive reserve be increased across the lifespan? And 3) does participation in different leisure and/or occupational activities in early life impact differently aging cognitive functioning? The studied sample of 349 participants from the Cleveland Longitudinal Aging Studies [107] has a mean age, years of education respectively of 74.8 and 15.9. The cognitive functioning in the study was assessed with the Modified Telephone Interview for Cognitive Status (TICS-m) [108] which is a telephone version of the MMSE. The results of the study obtained with path analysis [109, 110] suggest that reserve is dynamic but most amenable to change in early life, and educational pursuits in early life can positively impact cognitive functioning in late life. Lee et al. in [111] got interested in the type of social activity that reduces cognitive decline 4 years later among young-old (age 65-74) and old-old (age $\geq$ 75) adults. To that end, the authors used a sample of 1568 non-demented participants from two waves of the Korean Longitudinal Study of Aging [3]. The assessment of cognitive functioning is carried with the Korean version of the MMSE [112]. The authors used multiple linear regression analysis to evaluate the effects of 6 social activities on cognitive decline in the statistical analysis. The results of the study revealed that the participation in senior citizen clubs or having frequent contacts with adult children by phone or letters might help reduce cognitive decline in later life among older adults and the participation in various formal social activities may also have a beneficial

[2] Cognitive reserve is the mind's resistance to damage of the brain. Wikipedia

[3] https://survey.keis.or.kr/eng/klosa/klosa01.jsp

effect on preventing cognitive decline in older adults.

In a recent study, Rodrigues et al. [1] investigated the impact of the environment on cognitive health in older adults. The authors studied two hypotheses: the relationship between lifestyle factors and cognitive health in aging and the impact of stratification of the aging population to promote a better understanding of the aging process. The authors considered a sample of 3507 participants (average of age: 72 years; and an average number of years of education: 12.9) of the Health and Retirement Study (HRS) dataset[4] and used the ordinal logistic regression (OLR) [113] as a statistical model. As a dependent variable, the authors considered a composite word recall (immediate + delayed) [5] from which five cognitive categories (from 1 to 5 with 1 the lowest) were computed. In addition to the age, the number of years of education, and the sex of the participants, 35 lifestyle activities were selected in the model of analysis. The results of the research show that enriching lifestyle factors, particularly those that require extended periods of focus (ex. often reading and using a computer) have a positive relationship with cognitive health. In contrast, deleterious lifestyle activities such as financial constraints and smoking have a negative impact. Moreover, a marginal effect analysis[6] from the OLR shows a dichotomy pattern of the effects of the lifestyle factors on cognitive health among the sample of participants (see Figure 5.1). Indeed, while environmental factors such as housing problems, reading, and using computers have a high influence either positively or negatively on cognitive categories 1, 2, 4, and 5, the influence of these factors is low for cognitive category 3.

The results of the preceding study are similar to the orchid and dandelion theory observed within the children population in developmental sciences. We present a brief review of the theory in the next subsection.

5.1.2 Orchid and dandelion theory

In the 90s, a new theory referred to as orchid and dandelion theory [7] emerged in developmental sciences from the work of Thomas Boyce and Bruce Ellis

[4] https://hrs.isr.umich.edu/about

[5] Participants are asked to recall a list of 10 words either immediately or with a delayed time.

[6] A marginal effect analysis quantifies the amount of change in the dependent variable when a unit change of an independent variable is observed. In other words, it corresponds to the derivative of the dependent variable with respect to the independent variable.

[7] The terms dandelion and orchid are originated from flowers of the same names. The dandelion flowers are resistant flowers that can grow independently to their environmental conditions, whereas the orchid flowers' growth is highly related to their environmental conditions (either good or bad).

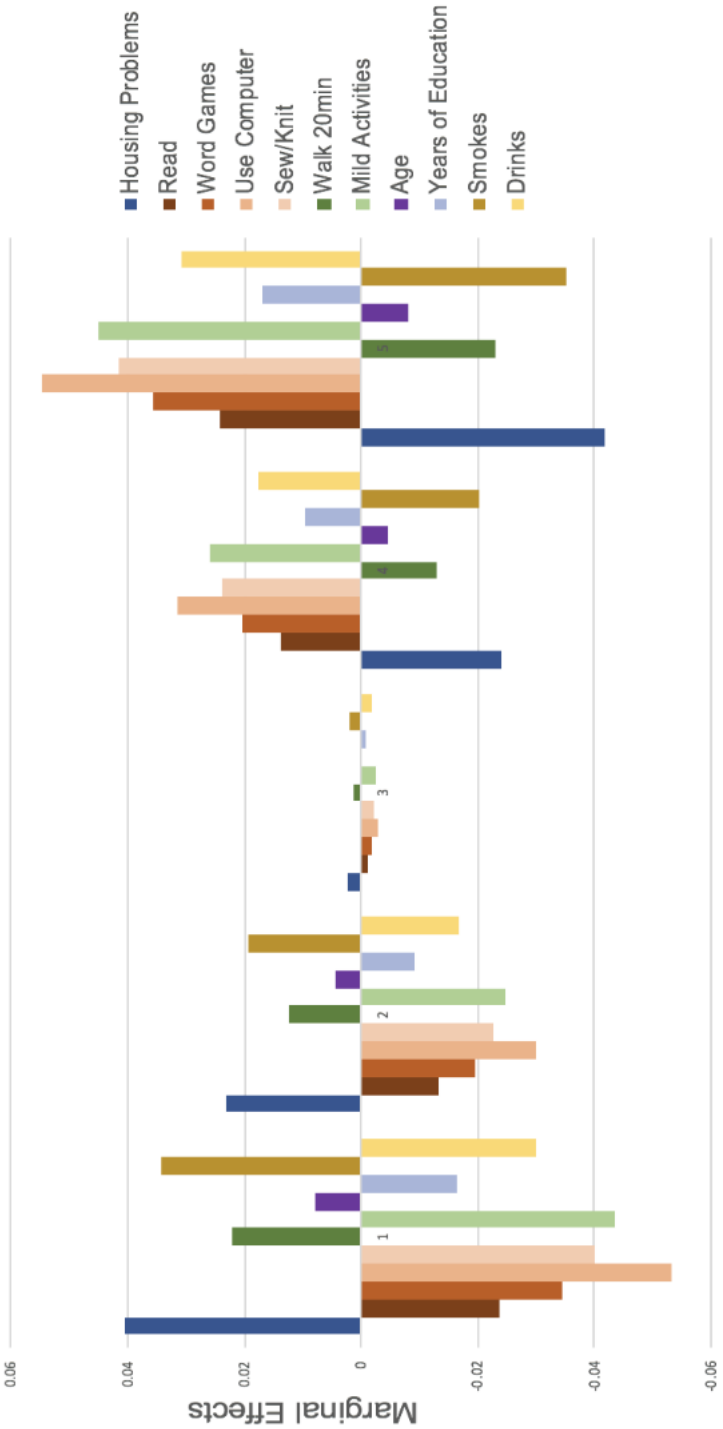

Figure 5.1: Marginal effects analysis from [1] on the most significant lifestyle factors from the OLR model (see subsection 5.2.1 for a description of each factor). The x-axis corresponds to the 5 cognitive categories and the y-axis to the marginal scores.

in [114]. In the latter paper, the authors examined the interactive influences of environmental stressors and biological stress reactivity on the incidence of respiratory illnesses in young children. Unexpected findings from the study show that the lowest rates of illness with the studied samples were found for equally high reactivity children reared in low-stress, highly supportive, and nurturant family environments. From the latter results, the authors introduced the orchid and dandelion theory which stated two different types of reactions to adversity within children: the sensitivity to environment children labelled as orchids respond positively and negatively under good or poor environmental conditions, whereas the dandelions children are less affected by environmental conditions. Over two decades, the theory has been developed, and different studies (clinical and data-driven) have been conducted to replicate the new observations and better understand the phenomenon. In [115], Belsky studied the effect of maternal sensitivity on infant attachment security and concluded that children vary in their susceptibility to rearing experience. In [116], the studies of Barr et al. on young rhesus macaques highlighted that in more stressful contexts, highly sensitive or orchid-like young monkeys are likely to do poorly, whereas, in supportive settings, the same individuals show better outcomes than their robust, relatively insensitive dandelion-like peers. Studies on the interactions between genetics and environment (GxE) revealed that some individuals and animals with particular genotypes are more susceptible than others [117]. In addition to the genetic and environmental factors, authors in [118] highlighted the necessity to consider the developmental time component in the context of biology adversity and resilience studies. While some longitudinal studies reported the fading of differential susceptibility over time (ex. see [119, 120]), contradictory results (increasing of differential susceptibility over time) have been published [121]. Nonetheless, both types of results suggest a monotone behavior of differential susceptibility over time.

While the preceding studies focused on the replication of the orchid and dandelion theory within children, any of the studies investigated the observation of the phenomenon within older adults despite the worldwide exponential increase of this proportion of the population. For the first time, the phenomenon has been observed within the older adult population in [1]. As for the orchid and dandelion phenomenon within the children population, validation of the findings in [1] is needed. To that end, we conduct a complimentary analysis by using the clustering methods *CFE* and *cat-ECM* as alternatives to the OLR model.

Due to the nature of medical and health data, special mathematical and computer science frameworks are needed to mine them. The next subsection presents the use of soft computing in medical and health researches.

5.1.3 Soft computing in medicine and healthcare

Health data sets, usually collected through longitudinal surveys [8] offer a good way for studying individual trajectories particularly older adults. In general, these data are collected through questionnaires and interviews. Depending on the nature of the studies, the characteristics of the data vary on the sample size, modules of questionnaires, and the types (ex. raw, images, medical and non-medical) of the collected data. In addition to the preceding specificity, health data incorporate uncertain and imperfect knowledge due to their nature [122]. To handle uncertainty and imperfection in these data several techniques such as soft computing[9] have been proposed in the literature. A recent review of papers published between 1991 and 2020 on handling uncertainty in medical data shows that the most frequent techniques used are Bayesian inference, fuzzy systems, Monte Carlo simulation, rough classification, Dempster-Shafer theory, and imprecise probability [123]. In [124], the authors surveyed the utilization of fuzzy technologies particularly fuzzy logic in different medical domains ranging from internal medicine to image and signal processing and biomedical laboratory tests. The same authors reported the development of a health status index of patients proposed in [125] as the first application of fuzzy logic in healthcare. The development of the index was motivated by the fuzzy boundaries of patient statuses. In addition to fuzzy logic, fuzzy clustering can be used to find relationships between patients and to assist physicians [125, 126]. As the Fuzzy Sets Theory, the Dempster-Shafer theory has also been used for managing uncertainty in health data sets. In [127], SShenoy discussed how the theory can be used in the framework of valuation-based systems that serve as a framework for managing uncertainty in expert systems. The Dempster-Shafer theory has been used in health data analysis such as heart and brain strokes prediction [128], medicine recommendation [129], breast cancer tumors prediction [130], obesity epidemic [131], prostate cancer prediction [132], medical information fusion [133, 134, 135], drug-drug interactions [136] and so on.

In the next section, we present the protocol and the results of the experiments.

5.2 Experiments

In this section, we propose a complementary analysis to investigate the replication of the findings from Rodriguez et al [1]. To that end, we use the same data set as the previous authors and use the developed clustering methods discussed in Chapter 3 and 4 as alternatives to the ordinal logistic regression

[8] The same individuals are followed up during several years.

[9] Ensemble of techniques for reasoning and modeling with uncertain, imprecise, and incomplete information.

model [137] in [1]. Therefore, our goals are 1) to verify whether or not we can obtain 5 clusters with each cluster corresponding to one cognitive category, and 2) to determine the effect of each lifestyle factor in the clusters as the marginal effect analysis in Figure 5.1. The interests of the new methods are two-fold. On one hand, contrary to supervised learning methods such as OLR, the new clustering methods *CFE* and *cat-ECM* do not require labeled data offering therefore more applicability of these methods compared to supervised learning methods. On the other hand, with the HRS data set, we are expecting both methods to capture fuzzy boundaries between orchids and dandelions older adults and the uncertainty in the assignments of the individuals in these classes and subsets of classes.

5.2.1 Data set

In this subsection, we provide a detailed description of the HRS data set.

The HRS data set is a longitudinal survey data started in 1992. The initial number of participants in the database was 41 000 over 50 years old. The data set has many components such as surveys on income and wealth, health and use of health services, employment, and psycho-social and lifestyle factors. In [1], the authors considered the data from the 2012 and 2016 waves. In order to have the same individuals in both waves and with no missing values, the data were filtered, and the final sample size is 3507. To model the impact of environmental factors (lifestyle factors) on cognitive health with the OLR model, a composite cognitive function score composed of the immediate and delayed word recall tests was considered as a dependent variable in [1]. A stratification of the final score (referred to as Cogn) values ranging from 0 to 20 (with 20 being the highest score) into five cognitive categories (from 1 to 5 with 1 corresponding to the lowest category) was applied in order to have a uniform distribution of participants in each category. The number of individuals in these categories is respectively 735, 786, 971, 868, and 616 respectively for cognitive categories 1 to 5. With the latter stratification, from the work of Rodrigues et al, cognitive categories 1, 2, 4, and 5 correspond to orchids and the 3rd category to dandelions. In our analysis, we will consider the top eleven [10] most significant factors from the OLR model in [1] to fit our models. The marginal effects of these factors correspond to Figure 5.1. Among the factors, the age of participants and their number of years of education is considered in [1] as control variables. The remaining seven variables were binarized in the latter study to indicate whether or not the participants are concerned by the cor-

[10] The choice of the most eleven significant factors holds on the fact that these factors have the most significant impact on cognitive health as it can be seen in Figure 5.1. Also, for a preliminary replication analysis, we were interested in running the analysis with few variables in order to reduce the noise in the data.

Variables	Descriptions
Housing Problems	Indicates if the participants have housing problems.
Read	Indicates if the participants often read.
Word Games	Indicates if the participants often do word games.
Use Computer	Indicates if the participants often us a computer.
Sew/Knit	Indicates if the participants often sew or knit.
Walk 20min	Indicates if the participants often walking 20 minutes.
Mild Activities	Indicates if the participants often do mild activities.
Smokes	Indicates if the participants smoke.
Drinks	Indicates if the participants often drink.

Table 5.1: Binary variables from the HRS dataset used in our cluster analysis.

responding lifestyle factors. The binary variables are described in Table 5.1.

Table 5.3 corresponds to the mean years of education and age in each cognitive category. From this table, the lowest cognitive category (1) corresponds to old-old adults in the sample with the lowest number of years of education whereas the highest cognitive category (5) corresponds to young-old adults with the highest number of years of education. We provide the descriptive statistics of the selected lifestyle factors for the cluster analysis in Table C.1 in Appendix C. The profiles of individuals in each cognitive category are described in Table 5.2.

To visualize the HRS data, we use principal components analysis and plot cognitive categories in Figure 5.2. In the latter figure, the top left, top right, bottom left and bottom right sub-figures respectively correspond to the plotting of all cognitive categories, cognitive categories 1 and 5, cognitive categories 2 and 4, and cognitive categories 1, 3, and 5. It can be noted that in the first sub-figure that there is an overlap between almost all cognitive categories. In the second sub-figure, the difference between the extreme cognitive categories (1 and 5) can be seen on the x-axis.

In the next subsection, we describe the evaluation criteria used in the experiments.

5.2.2 Evaluation criteria

To evaluate the performances of *CFE* and *cat-ECM* on the HRS data set and the similarities between the obtained clusters and the cognition (Cogn) variable we use the same internal and external evaluation criteria as in Chapter 3. Namely, we use the partition coefficient, the partition entropy, the rand index, the fuzzy rand index, and the fuzzy silhouette index. In order to ex-

Cognitive categories	Profiles
1	Most individuals often read and do not have housing problems. In contrast, they do not often do mild activities, do word games, use a computer, sew or knit, walk 20mins, drink, and smoke.
2	Most individuals often read, walk 20mins and drink, and do not have housing problems. In contrast, they do not often do mild activities, do word games, use a computer, sew or knit, and Smoke.
3	Most individuals often have mild activities, read, walk 20mins, use a computer, drink, and do not have housing problems. In contrast, they do not often do word games, sew or knit, and Smoke.
4	Same profile as in cognitive category.
5	Most individuals often have mild activities, read, do word games, walk 20mins, use a computer, drink, and do not have housing problems. In contrast, they do not often sew or knit and Smoke. mild activities and use computers

Table 5.2: Profiles of individuals in each cognitive category based on the attribute categories frequencies.

	Years of Education	Age
Cogn		
1	11.51	76.22
2	12.42	73.12
3	12.95	71.47
4	13.44	70.28
5	14.34	68.83

Table 5.3: Mean number of years of education and age per cognitive category.

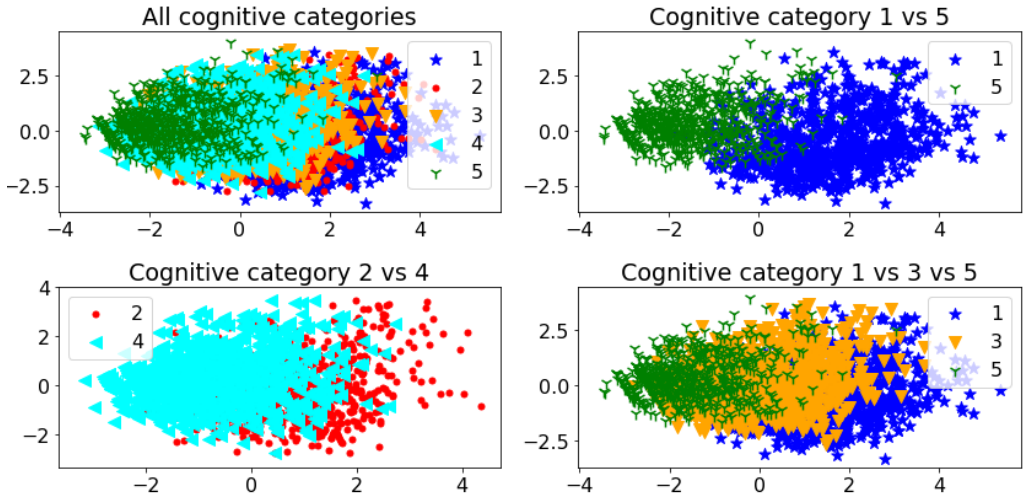

Figure 5.2: Projection of the HRS data with principal component analysis with different cognitive categories configurations. The x and y axes respectively correspond to the first and second components which explain 19% and 12% of the total inertia (variance).

amine the behavior of each lifestyle factor in the clusters produced by *CFE* and *cat-ECM*, we use the test value statistic [138] defined by Equation (5.1).

Given a sample of objects and their classes (obtained clusters from the algorithms in our case), the test value determines the significance of the variables in each cluster. For numerical variables, the test value compares the mean of each variable to the overall mean of the sample. Similarly, for categorical data, the test value compares the proportions of the corresponding categories of the variables in the clusters to the overall proportion of the category in the sample. As all the input variables of *CFE* and *cat-ECM* are categorical, we use the test value for categorical which is given by the following equation

Let a_l be the category for which we want to compute the test value in a cluster (e.g C). Let $p_{l/C}$ be the proportion of the category in C and n_C be the number of objects in C. Let p_l be the overall proportion of the category and n the total number of objects in the sample. The test value (referred to as vTest) is then defined by

$$vTest(a_l) = \sqrt{n_C} \times \frac{p_{l/C} - p_l}{\sqrt{\frac{n - n_C}{n - 1} \times p_l \times (1 - p_l)}}. \tag{5.1}$$

As it can be noted in Equation (5.1), the test value is similar to a z-score

and is approximately normally distributed [138]. Therefore, critical values for a two-sided significance test at 5% are −1.96 and 1.96. A negative (respectively positive) value of the test value indicates that the proportion of the corresponding category in the cluster is statistically lower (respectively greater) than the overall proportion of the sample. From this value, by considering category 1 for each lifestyle factor, we can determine the influence of the latter factors on cognitive health[11].

In the next subsections, we describe the protocol of the experiments, the materials used to conduct the experiments, and the results of the cluster analysis.

5.2.3 Experimental protocol

In the experiments, we run the *CFE* and *cat-ECM* algorithms over ten runs with different initialization and computed the mean scores and standard deviations. We use the test value to determine the statistical significance of each attribute category in the obtained clusters. Then, we compare the test values of the lifestyle factors categories in the clusters to the corresponding test values in each cognitive category.

5.2.4 Parameter settings

To run the experiments, we set the number of clusters of *CFE* and *cat-ECM* to the number of cognitive categories (5) and the values of parameters m and α of the *CFE* algorithm respectively to 1.1 and $1e^{-2}$. For *cat-ECM* we set β to 1.1, α to 1.4 and ρ to 6.

5.2.5 Materials

We use the same materials presented in Chapter 3 and 4 in the experiments.

5.2.6 Results

In this subsection, we present the mean scores of *CFE* and *cat-ECM* on the HRS data set followed by comparisons of clusters obtained from the two methods and the comparisons of the test values of the attribute categories with the corresponding test values in the five cognitive categories.

[11] As the factors considered in the cluster analysis are binary, the test value can be computed for both categories 0 and 1. The test values of one category are sufficient for determining the influence of the corresponding factor. Therefore, we choose to compute the test values associated with the category 1.

	PC	PE	RI	FRI	FS
CFE	0.87 (0.06)	0.19 (0.09)	0.67 (0)	0.65 (0.01)	0.26 (0.02)
cat-ECM	0.65 (0.07)	0.63 (0.07)	0.64 (0.02)	0.59 (0.02)	0.29 (0.04)

Table 5.4: Mean scores of *CFE* and *cat-ECM* on HRS data set. The values in brackets correspond to the standard deviations. For external measures the scores are computed between the cognition variable (Cogn) and the obtained clusters.

Comparisons of scores and partitions

Table 5.4 corresponds to the mean scores and standard deviations obtained with *CFE* and *cat-ECM*. From this table, *CFE* obtain the best performances except for the FS score. In addition to the score comparisons, we also compared the hard partitions obtained from *CFE* and *cat-ECM*. For this purpose, we compute the relative and absolute frequencies of each cognitive category in the clusters. Figure C.1 and C.2 in Appendix C correspond respectively to the frequencies obtained from *CFE* and *cat-ECM*. It can be first noted on the latter figures that all the clusters from the two methods contain the five cognitive categories. Second, we can note that based on the absolute frequencies of cognitive categories some clusters (1 and 3 for *CFE* and 1 and 2 for *cat-ECM*) contains mostly individuals with cognitive categories 1 and 2. Similarly, the clusters 2 and 5 for *CFE* and 3 and 5 for *cat-ECM* contain mostly individuals with cognitive categories 4 and 5. Third, in both of the two partitions, we can note that there is a cluster (cluster 4 for *CFE* and *cat-ECM*) in which the five cognitive categories are approximately uniformly distributed. Consequently, for *CFE* we merge the clusters 1 and 3 (referred to as C{13}, which contains 1466 objects) and the clusters 2 and 5 (referred as C{25} and contain 2086 objects). For *cat-ECM* we merged the clusters 1 and 2 (referred as C{12} and contain 1990 objects) and the clusters 3 and 5 (referred as C{35} and contain 1768 objects). We let the cluster 4 of the two methods which contains respectively 424 and 218 objects and referred it as C{4}.

The new relative and global percentage of cognitive categories are reported in Figure 5.3 and Figure 5.4 for respectively *CFE* and *cat-ECM*.

Remark 15. Merged clusters from *CFE* and *cat-ECM* are note C{...} (e.g C{25}) whereas the focal sets from *cat-ECM* are noted C25, C12 etc...

In the next subsection, we present the results of the statistical comparisons.

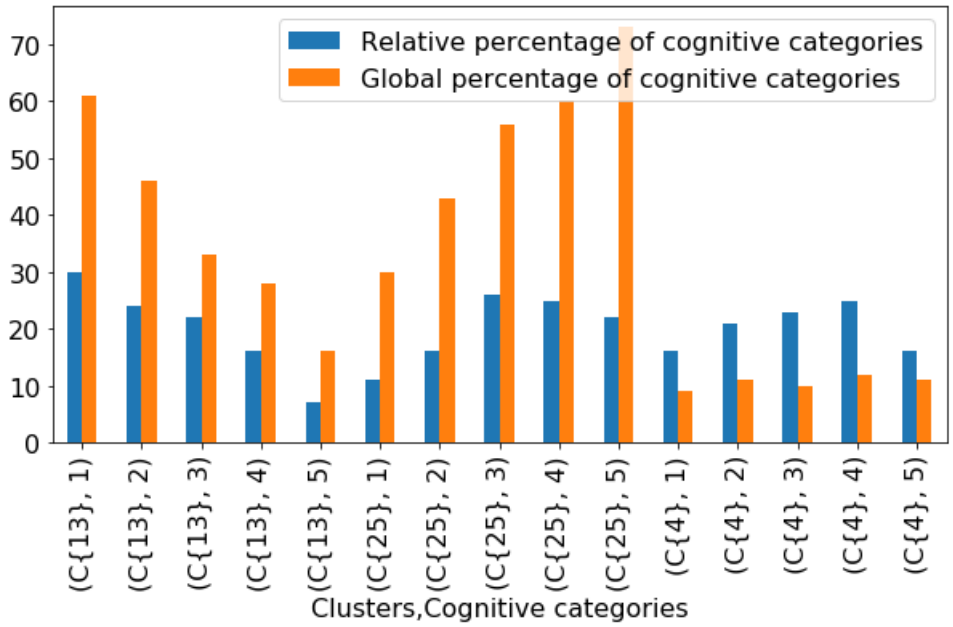

Figure 5.3: Relative and global percentage of cognitive categories in merged clusters from *CFE*.

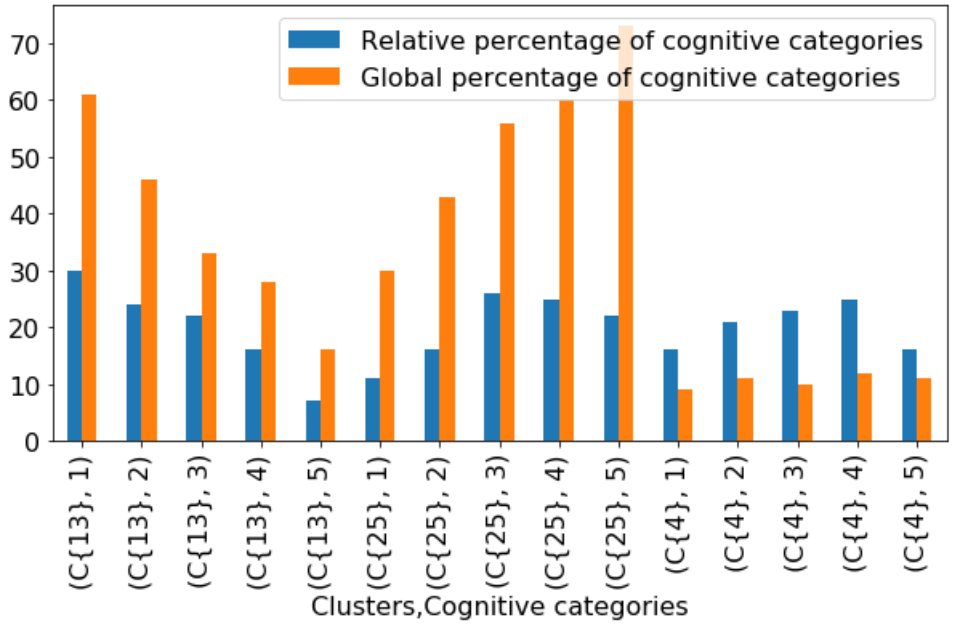

Figure 5.4: Relative and global percentage of cognitive categories in merged clusters from *cat-ECM*.

Comparisons of test values

We use the test value as an alternative to the analysis of the marginal effects in Figure 5.1 to determine the effect of each lifestyle factor on cognitive health. We first compute the test values of the lifestyles (category 1 for each lifestyle factor) in each cognitive level (see Figure C.3 in Appendix C). The latter test values correspond to the expected values when there is a 1-to-1 correspondence between cognitive categories and the clusters from *CFE* and *cat-ECM*. As we merge the clusters of orchid older adults with low and high cognitive categories, we also combine cognitive categories 1 and 2 and 4 and 5 in the cognition variable Cogn. The cognitive category 3, representing the dandelions is kept as is. The obtained test values are summarized in Figure 5.5. We can note in this figure a similar opposite effect of the lifestyle factors on cognition such as in the analysis of the marginal effects in Figure 5.1. Indeed, while the test values of most of the lifestyle factors categories are significant for individuals with low and high cognitive categories, only a few lifestyles are significant (Use Computer 2.9 and Drink 2.5) for individuals with cognitive category 3.

As a quick reminder, a positive (respectively negative) test value of a lifestyle factor category indicates that the individuals in the corresponding class have a significantly high (respectively low) percentage of that category compared to the overall category frequency. Therefore, in cognitive class {12} the individuals do not often do mild activities, do not often read, do not often do word games, do not often use a computer, do not often sew or knit, do not often walk 20 min, have housing problems, do not often drink and often smoke. These results are overall consistent with the marginal effect analysis. We can note that the most significant lifestyle factor is Use Computer which is also consistent with the marginal effect analysis.

Figures 5.6 and 5.7 correspond respectively to the test values of the merged clusters from *CFE* and *cat-ECM*. It can be noted that there is an opposite behavior of the lifestyle factors in the classes with respectively low and high cognitive category. We can also notice that the test values in cluster C{4} from *CFE* and *cat-ECM* are overall lower than those of the preceding two classes. In the latter cluster, for *CFE* the lifestyles factors Use Computer and Housing problems are not significant (their test values are respectively -1.77 and -0.48), for *cat-ECM* the test value is not significant only for lifestyle factor Sew/Knit. From these observations, the test values of merged clusters from *CFE* and *cat-ECM* that contain orchids of older adults are consistent with the analysis of the marginal effects. For clusters C{4} from the two methods that contain uniform distribution of the five cognitive categories, the test values should be interpreted with cautions as explained few paragraphs later.

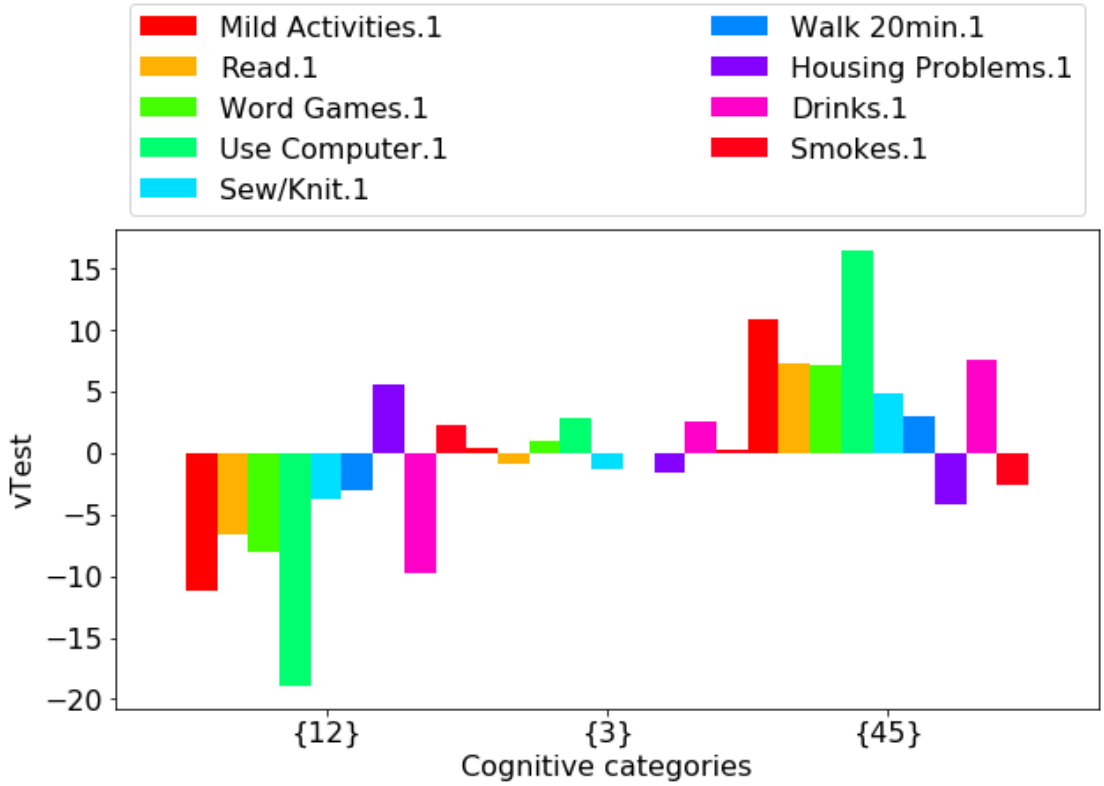

Figure 5.5: Test values of lifestyle factors categories in each cognitive category where the classes {12} and {45} correspond respectively to the merge of cognitive categories 1 and 2 and 4 and 5. The class {3} corresponds to cognitive category 3.

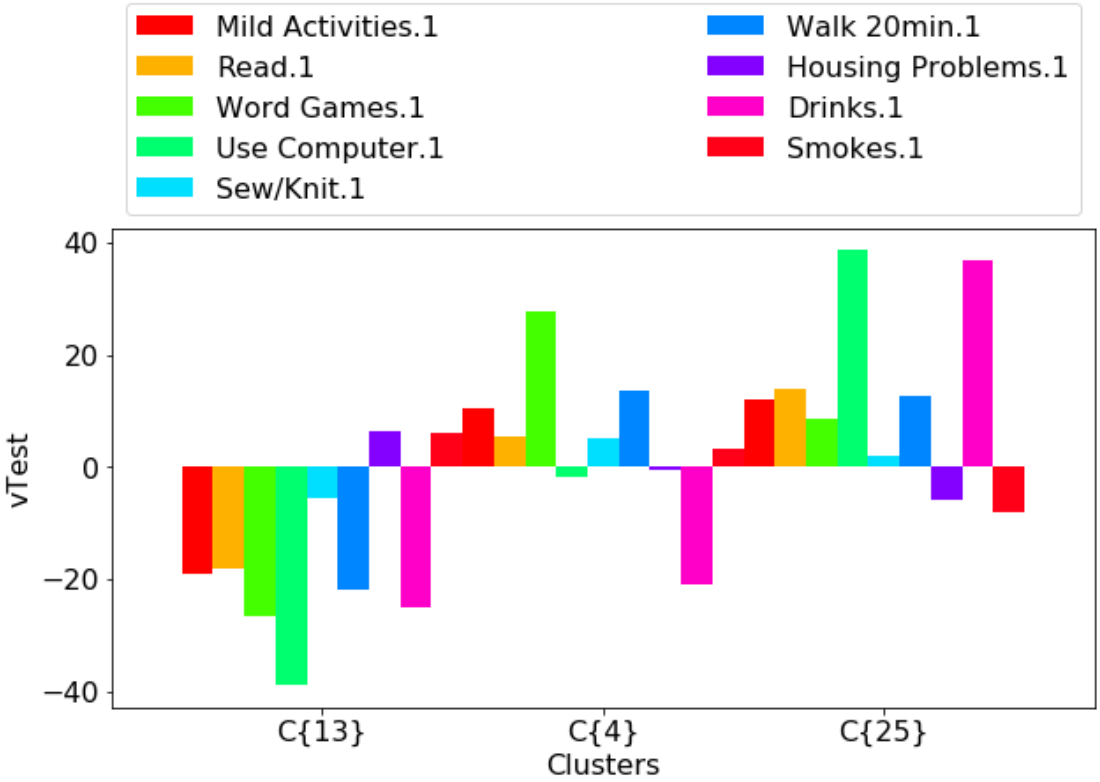

Figure 5.6: test values of lifestyle factors categories obtained from the hard partition of *CFE*. The classes {14} and {235} correspond respectively to the merging of clusters 1, 4 and 2, 3, 5 and correspond to older adults respectively with low and high cognitive levels.

To take advantage of the evidential partition generated from *cat-ECM* with the number of clusters equals 5, we compute the hard partition of the focal sets by assigning the objects to the clusters with the highest membership degree. We then compute the test values of each lifestyle factor. Figure 5.8 describes the distribution of cognitive categories in each focal set. In the latter figure, the bars correspond to the number of objects in cognitive and focal set categories.

It can be noticed in this figure that focal sets C1 to C5 contain respectively mostly cognitive categories 3, 4, 1, and 2. For Ω, the most frequent cognitive category is 3. We can also note that subsets with cardinality two successfully retrieve individuals in corresponding singletons. For instance, the subset C12 contains mostly individuals with cognitive categories 3 and 4 which correspond to the highest cognitive categories respectively in sin-

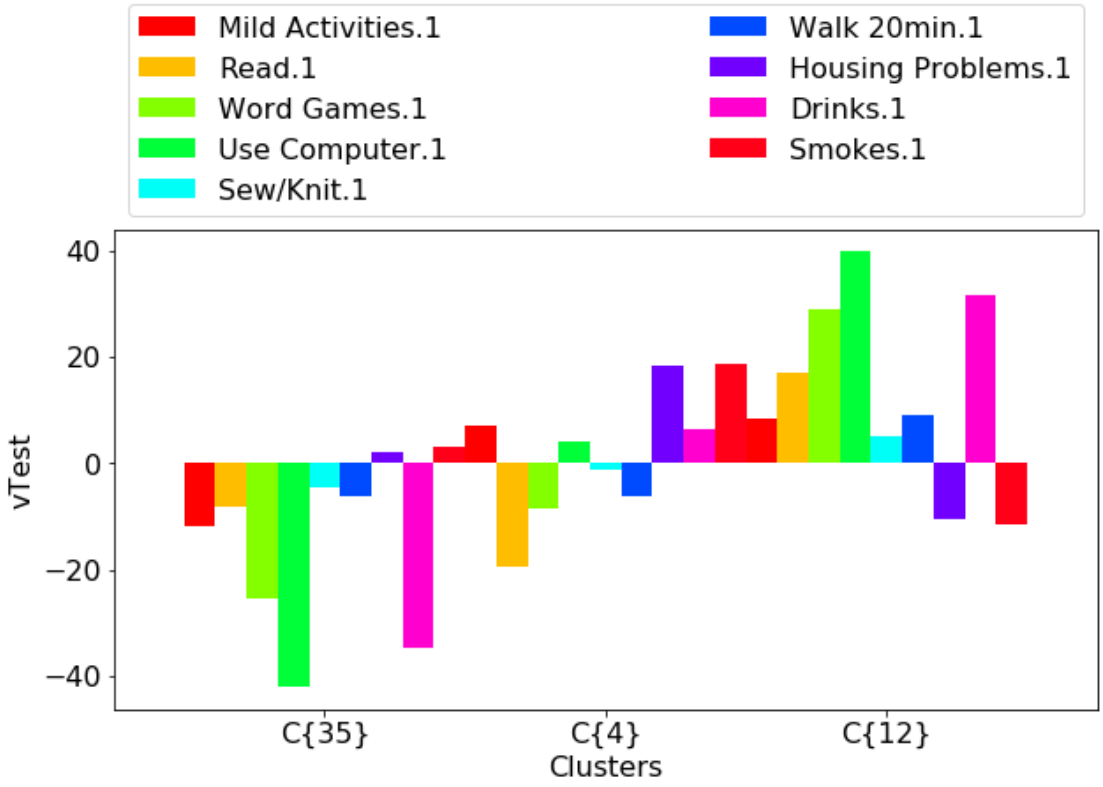

Figure 5.7: test values of lifestyle factors categories obtained from the hard partition of *cat-ECM*. The classes {12} and {345} correspond respectively to the merging of clusters 1, 2 and 3, 4, 5 and correspond to older adults respectively with high and low cognitive levels.

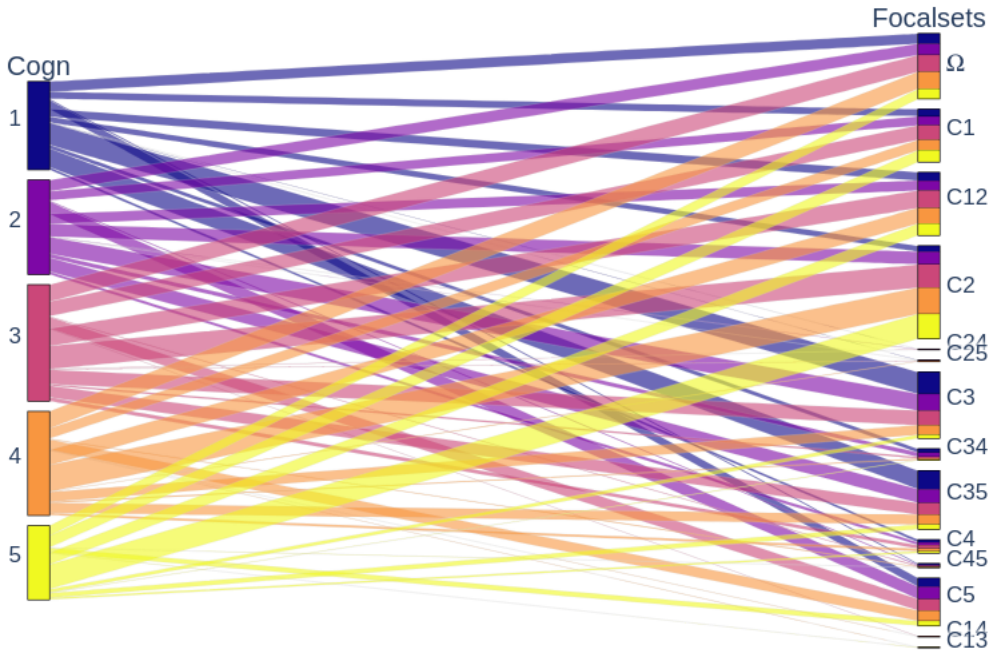

Figure 5.8: Distribution of cognitive categories in focal sets obtained from *cat-ECM*.

gletons C1 and C2. Figure 5.9 corresponds to the Venn diagram of the focal sets which shows the ability of *cat-ECM* to capture uncertainty in object assignments to clusters. Indeed, individuals in subsets with cardinality 2 are expected from cluster assignments as it is known in the literature that in addition to individuals with high sensitivity (orchids) and low sensitivity (dandelions) there also exists individuals with medium sensitivity referred to as tulips [139].

We compute the test values of each lifestyle factor of the focal sets from *cat-ECM*. The results are reported in Figure 5.10. We can note in that figure that the test values of subsets C13 (2 objects), C14 (3 objects), C24 (3 objects), and C25 (10 objects) are the lowest. This observation is due to the sensitivity of the test value to the sample size. Indeed, as it can be noticed in Equation 5.1, there is a factor $\sqrt{n_C}$ which corresponds to the number of objects in the corresponding cluster the test value is computed. Hence, when the number of objects is multiplied (respectively divided) by 100 the value of the test is multiplied (respectively divided) by 10. In Figures 5.6 and 5.7, there is a factor about 2 and 3 between the merged clusters of *CFE* and *cat-ECM* and the clusters C{4}.

In the next section, we discuss the results of the statistical analysis.

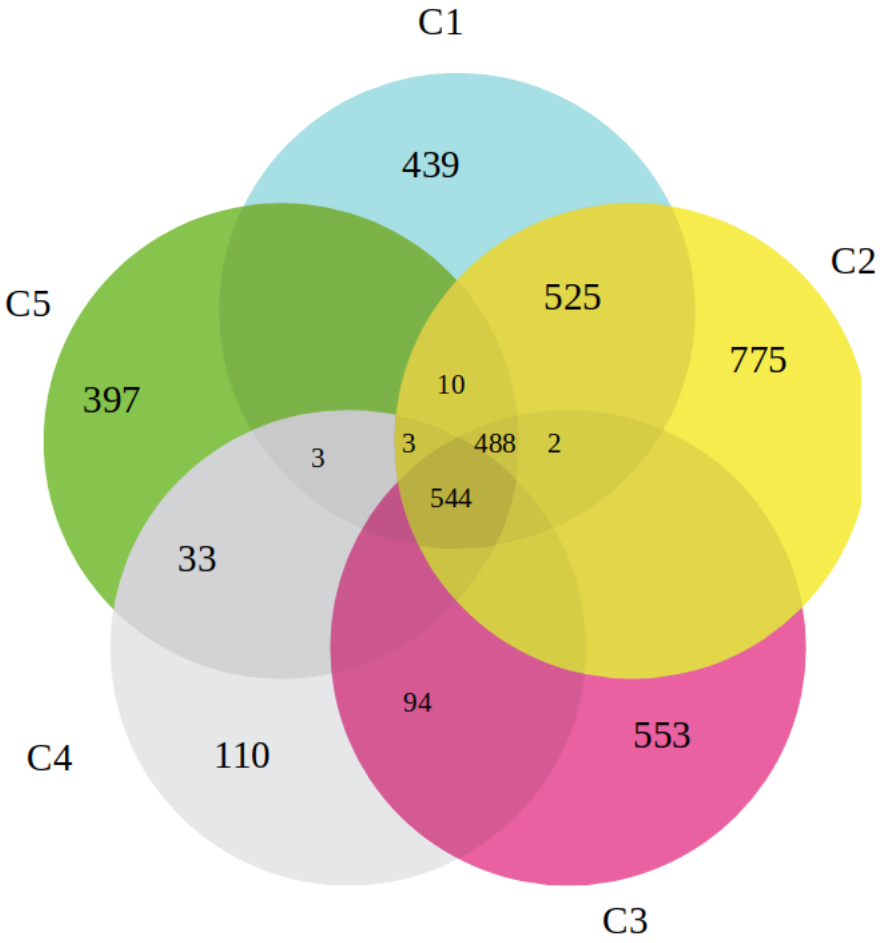

Figure 5.9: Venn diagram from the focal sets obtained from *cat-ECM*. The numbers correspond to the number of objects in each subset.

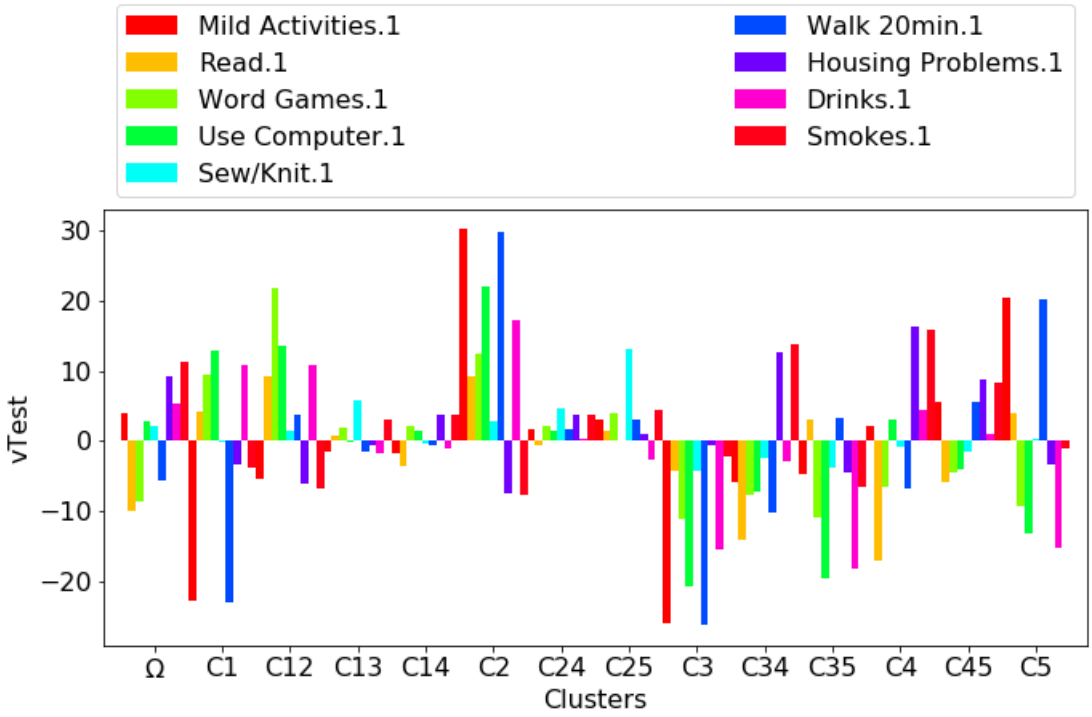

Figure 5.10: Test values of focal sets obtained from *cat-ECM*.

5.3 Discussions

The statistical analysis results described in the previous section show that *CFE* and *cat-ECM* are able to successfully capture older adults in the HRS data set with similar cognitive categories. While orchid older adults are identified by the two methods in different clusters, most of the dandelions older adults are assigned into the merged clusters of *CFE* and *cat-ECM*, and the clusters C{4} in which there is a uniform distribution of all cognitive categories. The latter distribution can be explained by the fact that *CFE* and *cat-ECM* are frequency-based clustering methods. Indeed, as described in Table C.1 in Appendix C, the most frequent categories of the lifestyle factors in cognitive levels 2 and 3 on one hand and 3, 4, 5 in another hand are similar. As *CFE* and *cat-ECM* cluster centers are based on the frequencies of the attributes categories, it is then not surprising to observe the distribution of individuals with cognitive level 3 in other clusters.

Despite the sensitivity of the test value to the sample size, the results of our analysis show promising results as we use unsupervised learning compared to the ordinal logistic regression used in [1]. Indeed, with cluster analysis even with desired output classes such as in our case (cognitive categories), there is no guarantee that the obtained clusters will correspond or be similar to the output classes. In addition, supervised learning methods such as the ordinal logistic regression are able to capture non-linear relationships between dependent and independent variables in particular in health data sets that are known to be mostly non-linear [140]. Therefore, *CFE* and *cat-ECM* which are linear methods show encouraging results.

5.4 Conclusion

In this chapter, we investigate the replication of new findings on the impact of environmental factors on the cognitive health of older adults. Our results show that older adults with low and high cognitive levels that behave positively when the conditions are positive and negatively when the latter are negative represent the environment-sensitive individual - orchids. Older adults with mild cognitive levels that are less sensitive to environmental conditions represent environment non-sensitive individual - dandelions. In our process of replication, we use the most significant lifestyle factors from [1] and use cluster analysis with *CFE* and *cat-ECM* as an alternative to the logistic regression-based model used by the authors. We first compare the scores of the internal and external measures of PC, PE, RI, FRI, and FS of *CFE* and *cat-ECM*. Based on these scores, *CFE* produced better performances compared to *cat-ECM*. In the next step, we compare the hard partitions of the two methods with the cognitive variable. From the latter

comparisons, we merge some clusters of the two methods and compute the statistical test values which compare the proportions of lifestyle factors categories in the obtained clusters to their overall proportions in the sample. From the results of this analysis, we determine the most significant lifestyle factors categories in each cluster which show opposite interactions between lifestyle factors and cognitive health similar to the original findings. On the overall, *CFE* and *cat-ECM* capture all cognitive categories. While the orchid's older adults are approximately grouped as expected, the dandelions were distributed among clusters. The latter observation can be explained by the fact the frequencies of the lifestyle factors categories or dandelions are similar to the orchids and as *CFE* and *cat-ECM* are frequency-based methods. Nonetheless, our methods were able to capture imprecision and uncertainty of object assignments into the clusters. Finally, the obtained results show new research directions of the replication of orchid and dandelion phenomenon in older adults with cluster analysis.

Improvements of our experimental results can consist in first, selecting a statistical measure less sensitive to the sample size such as Cohen effect size [141]. Second, the two clustering methods *CFE* and *cat-ECM* can be applied to the original lifestyle factors without binarization used in [1]. Third, different parameter settings of the clustering methods can be used as we limit our analysis to several chosen parameters.

Key points

- The analysis of the dataset HRS in [1] led to the observation of the orchid and dandelion phenomenon in older adults. We conduct a cluster analysis to investigate the replication of the findings.
- We use *CFE* and *cat-ECM* in the cluster analysis. The statistical test values computed from the obtained clusters show opposite behavior of the lifestyle factors on orchids older adults with low and high cognitive categories. The two methods distributed the dandelions older adults into other clusters with one of the clusters containing a uniform distribution of all cognitive categories.
- Our analysis shows encouraging results despite 1) the sensitivity of the test value statistic to the number of objects in clusters, and 2) the linearity of *CFE* and *cat-ECM* in contrast to the non-linear method used in [1].

Publications and Communications

The results of cluster analysis have been presented in several laboratory meetings. We are planning to publish these results in an international journal.

Conclusion and perspectives

Conclusion

In this report, we propose two new clustering methods for categorical data. The first method referred to as *CFE*, uses the fuzzy sets theory to model the imprecision of object assignments to clusters and cluster centers. The second method referred to as *cat-ECM* uses the Dempster-Shafer theory of evidence to model the uncertainty of object labeling.

We introduce the *CFE* algorithm, an extension of the fuzzy k-modes clustering method. The objective function of *CFE* incorporates Shannon's entropy as a regularization function of the weights of attribute categories which indicate their importance. We conduct several experiments on nine data sets with different characteristics to compare the performance of *CFE* against existing numerical and categorical clustering methods. The comparisons are based on five evaluation criteria with both internal and external measures. In addition to these measures, we perform statistical comparisons based on critical difference diagrams. The results of the comparisons show that overall there is not a statistical significance between the methods. Nonetheless, *CFE* achieved good performance and sometimes even better than numerical clustering methods such as fuzzy c-means. As many clustering methods, *CFE* presents some limitations that we discussed and some solutions have been proposed to handle these issues. The main strengths of *CFE* hold on the ability to capture fuzzy boundaries in categorical data and the fuzzy representation of the centers of the clusters that can help to have more interpretability of the results.

We present the *cat-ECM* clustering algorithm which is an extension of the numerical *evidential c-means* (*ECM*) algorithm. The new method has the same objective function as *ECM*. Despite the use of fuzzy sets theory for cluster centers representation, the updating formula of singletons obtained through an alternate optimization scheme are hard. In the experiments, we use the same data sets and evaluation criteria as for *CFE*. The performance comparisons of the methods over the nine data sets show that overall there

is no statistical difference. Moreover, we compare the evidential partitions of *cat-ECM* and ECM based on the consistency with the true classes of objects and the nonspecificity. The results of these comparisons with some parameter settings highlight that, the partitions obtained from *cat-ECM* are preferable to those obtained with *ECM*.

Finally, we present a real-world application of *CFE* and *cat-ECM* in which we investigate the replication of new findings on the influence of environmental conditions on the cognitive health of older adults. To that end, we use the most significant lifestyle factors from the study [1] and use the two new methods as alternatives to the ordinal logistic regression model that leads to the findings. The test values statistics are alternative to the analysis of the marginal effects used by the authors. The results of the analysis show that *CFE* and *cat-ECM* were able to successfully capture older adults with high sensitivity to environment labeled orchids. Similar to orchid older adults, the methods were also able to capture dandelion older adults who have low sensitivity to environmental conditions. Due to the use of attribute category frequency similarities among the two types of individuals, the dandelion's older adults were uniformly distributed in the clusters containing orchids with low and high cognitive levels.

Perspectives

Several research directions can be proposed to improve the performances of the new methods and to develop new clustering methods:

- The use of non-linear dissimilarity measures into the objective functions of *CFE* and *cat-ECM*. As reported in [140], most health data sets features have non-linear relationships. Therefore, a modification of the objective function such as a non-linear distance can be used to capture these relationships. Other alternatives such as density, kernel, graph, and manifold fuzzy clustering approach [142] can be investigated.

- *CFE* and *cat-ECM* can be extended to fit time series data by using for instance categorical time series dissimilarity measures such as in [143]. These extensions can for example be used on all the waves of the HRS dataset in order to capture the dynamic evolution of the lifestyle factors and cognitive health.

- New clustering methods that can handle both numerical and categorical data at the same time can be derived by combining *CFE* and *FCM* in one hand and *cat-ECM* and *ECM* in another hand.

- The updating formula of the centers of singletons of *cat-ECM* is currently hard. To take advantage of fuzzy centers, penalization functions

can be used on the weights of attribute categories such as Shannon's entropy used in *CFE* and the L2 penalization defined as follows:

$$||W||_2 = \sum_{k=1}^{c} \sum_{l=1}^{p} \sum_{t=1}^{n_l} (w_{kl}^{(t)})^2, \tag{5.2}$$

where W is the matrix of attributes categories weights, c, p and n_l are respectively the number of clusters, the number of attributes and the number of categories per attribute.

- Feature selection process can be incorporated into the objective functions of the developed clustering methods by associating a weight to each attribute. The latter corresponds to the importance of the attribute and can be optimized by solving the optimization problems associated to the methods. In high dimensional data such as Health data, the feature selection can be used to determine the most relevant features in the data.

- In order to match the attribute categories frequencies in the data sets to the weights obtained by *CFE*, frequency-based dissimilarity measures of categorical data can be used (see [144]).

- With the fuzzy representations of the centers and their ability to model imperfect data, applications of *CFE* and *cat-ECM* can be considered in explainable artificial intelligence to explain classifiers.

Two Ph.D. theses are planned to continue the work of this thesis. In addition, masters internships and school projects are expected.

Appendix

A — Complementary experiments results on CFE

A.1 CFE scores

This Section contains the scores of *CFE* with different values of β on the datasets used in the experiments in Chapter 3. Optimal scores are in bold.

A.1.1 Rand Index (RI)

	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Soybean	**1**	**1**	**1**	**1**	**1**	0.93	0.95	0.97	0.95	0.95
Zoo	0.93	0.92	0.88	0.89	**0.98**	0.88	0.87	0.87	0.94	0.82
Breast	0.62	0.62	0.58	0.61	0.56	0.58	0.56	0.54	0.57	**0.63**
Lung	**0.64**	**0.64**	0.62	0.61	**0.64**	0.58	0.62	0.6	0.61	0.59
Credits	0.67	0.67	0.67	0.67	**0.68**	0.67	**0.68**	0.67	0.67	0.67
Votes	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76
Cars	0.49	0.51	0.51	0.5	0.48	0.51	0.51	0.49	**0.52**	0.48
Dermatology	**0.86**	0.77	0.76	0.76	0.78	0.74	0.68	0.59	0.64	0.22
Mushrooms	0.51	0.51	0.52	**0.53**	0.52	0.52	**0.53**	**0.53**	**0.53**	0.52
Mean	0.72	0.71	0.7	0.7	0.71	0.69	0.68	0.67	0.69	0.63
Standard deviation	0.18	0.17	0.17	0.17	0.19	0.15	0.15	0.17	0.16	0.21

Table A.1: RI scores obtained with *CFE* with different values of β on the datasets.

A.1.2 Fuzzy Rand Index (FRI)

	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Soybean	**1.0**	0.98	0.95	0.92	0.9	0.85	0.83	0.81	0.79	0.77
Zoo	0.93	0.92	0.87	0.86	**0.94**	0.85	0.82	0.81	0.85	0.67
Breast	**0.53**	0.52	0.51	0.51	0.51	0.51	0.51	0.51	0.5	0.51
Lung	**0.64**	0.62	0.58	0.57	0.56	0.53	0.53	0.53	0.52	0.52
Credits	**0.66**	0.65	0.64	0.62	0.62	0.61	0.60	0.58	0.57	0.57
Votes	**0.7**	**0.7**	0.69	0.69	0.68	0.67	0.67	0.66	0.65	0.64
Cars	0.48	0.49	**0.5**	0.49	0.48	**0.5**	0.49	**0.5**	**0.5**	**0.5**
Dermatology	**0.84**	0.7	0.65	0.63	0.61	0.54	0.51	0.5	0.5	0.5
Mushrooms	0.58	0.59	0.6	**0.62**	0.6	0.6	0.59	0.58	0.58	0.57
Mean	0.71	0.69	0.67	0.66	0.66	0.63	0.62	0.61	0.61	0.58
Standard deviation	0.18	0.17	0.15	0.15	0.16	0.14	0.13	0.12	0.13	0.09

Table A.2: FRI scores obtained with *CFE* with different values of β on the datasets.

A.1.3 Fuzzy Silhouette Index (FS)

	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Soybean	0.47	0.48	0.49	0.49	**0.5**	0.47	0.48	0.49	0.48	0.48
Zoo	0.47	0.50	0.51	0.49	0.61	0.50	0.51	0.51	**0.65**	0.46
Breast	0.16	0.17	0.18	0.18	0.19	0.19	0.2	0.2	0.21	**0.29**
Lung	0.12	0.12	0.11	0.12	0.13	0.12	0.14	0.13	**0.15**	0.13
Credits	0.23	0.24	0.25	0.26	**0.29**	0.28	**0.29**	0.27	0.27	0.28
Votes	0.54	0.55	0.56	0.57	0.57	0.58	0.59	0.59	**0.6**	**0.6**
Cars	0.11	0.12	0.12	0.12	0.13	0.13	**0.17**	0.15	0.15	0.14
Dermatology	0.20	0.17	0.18	0.19	0.2	0.18	0.14	0.11	0.15	**0.25**
Mushrooms	0.23	0.24	0.25	0.26	0.26	0.27	**0.28**	**0.28**	**0.28**	**0.28**
Mean	0.28	0.29	0.29	0.30	0.32	0.30	0.31	0.3	0.33	0.32
Standard deviation	0.17	0.17	0.18	0.17	0.19	0.17	0.17	0.18	0.2	0.16

Table A.3: FS scores obtained with *CFE* with different values of β on the datasets.

A.1.4 Partition Coefficient (PC)

	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Soybean	**1.0**	0.98	0.94	0.90	0.85	0.78	0.72	0.67	0.62	0.58
Zoo	**0.99**	0.96	0.86	0.86	0.87	0.80	0.72	0.69	0.65	0.35
Breast	**0.77**	0.73	0.70	0.67	0.64	0.63	0.61	0.59	0.59	0.58
Lung	**0.96**	0.89	0.74	0.63	0.55	0.48	0.43	0.4	0.38	0.37
Credits	**0.99**	0.93	0.85	0.80	0.76	0.73	0.7	0.68	0.65	0.64
Votes	**0.98**	0.96	0.95	0.93	0.92	0.90	0.88	0.87	0.85	0.83
Cars	**0.82**	0.75	0.65	0.57	0.50	0.44	0.33	0.26	0.26	0.26
Dermatology	**0.91**	0.64	0.53	0.46	0.41	0.27	0.25	0.25	0.25	0.25
Mushrooms	**0.96**	0.90	0.83	0.78	0.73	0.70	0.67	0.64	0.63	0.60
Mean	0.93	0.86	0.78	0.73	0.69	0.64	0.59	0.56	0.54	0.50
Standard deviation	0.08	0.12	0.14	0.16	0.18	0.20	0.21	0.21	0.20	0.20

Table A.4: PC scores obtained with *CFE* with different values of β on the datasets.

A.1.5 Partition Entropy (PE)

	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Soybean	**0**	0.04	0.11	0.21	0.32	0.45	0.56	0.65	0.75	0.83
Zoo	**0.02**	0.08	0.25	0.30	0.29	0.44	0.61	0.69	0.81	1.48
Breast	**0.34**	0.40	0.44	0.49	0.52	0.54	0.56	0.58	0.59	0.6
Lung	**0.07**	0.21	0.48	0.66	0.77	0.88	0.96	0.99	1.02	1.04
Credits	**0.02**	0.14	0.25	0.32	0.38	0.42	0.46	0.49	0.51	0.54
Votes	**0.03**	0.06	0.08	0.11	0.14	0.16	0.19	0.22	0.25	0.28
Cars	**0.28**	0.44	0.64	0.80	0.93	1.04	1.24	1.36	1.37	1.38
Dermatology	**0.15**	0.64	0.84	0.97	1.07	1.34	1.38	1.39	1.39	1.39
Mushrooms	**0.07**	0.18	0.28	0.35	0.41	0.46	0.50	0.54	0.55	0.58
Mean	0.11	0.24	0.37	0.47	0.54	0.64	0.72	0.77	0.8	0.9
Standard deviation	0.12	0.21	0.25	0.29	0.32	0.37	0.39	0.4	0.39	0.44

Table A.5: PE scores obtained with *CFE* with different values of β on the datasets.

A.2 Critical difference diagrams

This section contains complement figures of the critical difference diagrams and some data used in the comparisons.

A.2.1 FRI and RI scores of CFE, FC* and FCM on the datasets

Table A.6: FRI scores obtained with *CFE*, *FC**, and *FCM* for $\beta = 1.9$ on the datasets. This table correspond to the data for the critical difference diagram in Figure 3.5.

	CFE	FC*	FCM
Soybean	0.79	0.79	0.68
Zoo	0.85	0.8	0.73
Breast	0.5	0.5	0.57
Lung	0.52	0.53	0.5
Credits	0.57	0.57	0.5
Votes	0.65	0.66	0.62
Cars	0.5	0.49	0.5
Dermatology	0.5	0.51	0.5
Mushrooms	0.58	0.57	0.5

Table A.7: RI scores obtained with *CFE*, *FC**, and *FCM* for $\beta = 2$ on the datasets. This table correspond to the data for the critical difference diagram in Figure 3.6.

	CFE	FC*	FCM
Soybean	0.95	0.85	1
Zoo	0.82	0.87	0.89
Breast	0.63	0.51	0.87
Lung	0.59	0.59	0.67
Credits	0.67	0.67	0.56
Votes	0.76	0.76	0.79
Cars	0.48	0.48	0.54
Dermatology	0.22	0.56	0.66
Mushrooms	0.52	0.57	0.77

A.2.2 Partition Coefficient critical difference diagrams

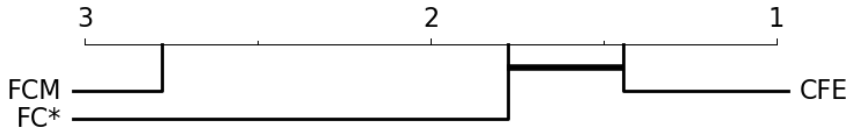

Figure A.1: Critical difference diagram obtained for PC scores with γ = 0.1 and β = 1.6.

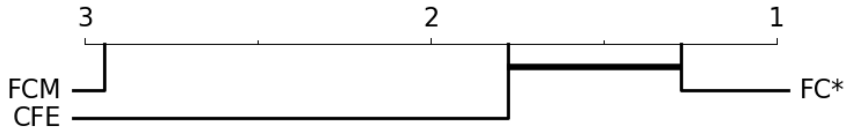

Figure A.2: Critical difference diagram obtained for PC scores with γ = 0.1 and β = 1.7.

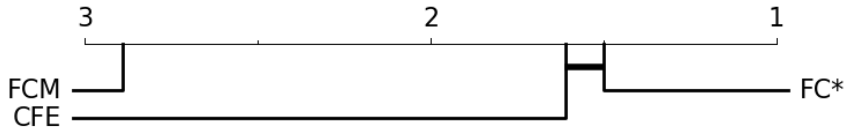

Figure A.3: Critical difference diagram obtained for PC scores with γ = 0.1 and β = 1.8.

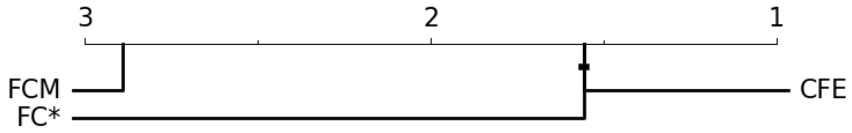

Figure A.4: Critical difference diagram obtained for PC scores with γ = 0.1 and β = 1.9.

A.2.3 Partition Entropy critical difference diagrams

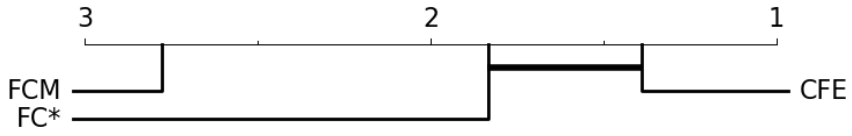

Figure A.5: Critical difference diagram obtained for PE scores with γ = 0.1 and $\beta = 1.6$.

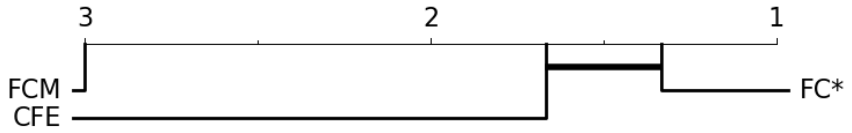

Figure A.6: Critical difference diagram obtained for PE scores with γ = 0.1 and $\beta = 1.7$.

A.2.4 Fuzzy Silhouette Index critical difference diagrams

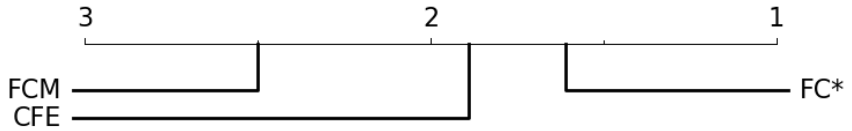

Figure A.7: Critical difference diagram obtained for FS scores with $\gamma = 0.7$ and $\beta = 1.8$.

B — Complementary experiments results on cat-ECM

B.1 cat-ECM scores

In this section, scores of *Ccat-ECM* with different values of β on the nine datasets are provided. The optimal scores are in bold.

B.1.1 Rand Index (RI)

	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Soybean	0.84	**1.0**	**1.0**	0.91	0.89	**1.0**	**1.0**	0.83	0.81	0.82
Zoo	0.88	0.94	0.93	0.95	0.86	0.92	0.91	0.89	**0.96**	0.86
Breast	0.51	**0.52**	**0.52**	0.51	0.51	0.51	**0.52**	**0.52**	0.51	0.51
Lung	0.62	**0.63**	0.62	0.59	0.32	0.48	0.52	0.6	0.6	0.61
Credits	0.67	0.67	**0.71**	**0.71**	0.67	**0.71**	**0.71**	**0.71**	**0.71**	**0.71**
Votes	0.76	**0.79**	**0.79**	**0.79**	0.76	**0.79**	**0.79**	0.76	**0.79**	**0.79**
Cars	0.48	**0.49**	0.48	**0.49**	0.48	0.48	**0.49**	0.47	0.48	**0.49**
Dermatology	**0.79**	0.72	0.74	0.74	0.69	0.66	0.54	0.67	0.52	0.6
Mushrooms	0.67	0.55	0.57	**0.69**	0.63	0.63	0.63	0.63	0.63	0.63
Mean	0.69	0.7	0.71	0.71	0.65	0.69	0.68	0.68	0.67	0.67
Standard deviation	0.14	0.18	0.18	0.16	0.18	0.19	0.19	0.14	0.16	0.13

Table B.1: RI scores obtained with *cat-ECM* with different values of β on the datasets.

B.1.2 Fuzzy Rand Index (FRI)

	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Soybean	0.74	**0.94**	0.9	0.84	0.79	0.78	0.72	0.70	0.67	0.67
Zoo	0.86	0.89	**0.92**	0.91	0.82	0.8	0.78	0.77	0.76	0.69
Breast	**0.52**	0.51	0.51	0.51	0.5	0.5	0.5	0.50	0.50	0.5
Lung	**0.58**	0.55	0.54	0.52	0.5	0.5	0.5	0.51	0.51	0.51
Credits	**0.65**	0.63	0.63	0.61	0.59	0.58	0.57	0.56	0.56	0.55
Votes	**0.71**	0.7	0.69	0.68	0.67	0.66	0.65	0.64	0.63	0.63
Cars	0.47	0.49	0.49	0.49	**0.5**	**0.5**	**0.5**	0.49	0.49	**0.5**
Dermatology	**0.73**	0.57	0.55	0.54	0.52	0.51	0.51	0.51	0.50	0.51
Mushrooms	**0.94**	0.66	0.65	0.64	0.62	0.61	0.59	0.58	0.58	0.57
Mean	0.69	0.66	0.65	0.64	0.61	0.6	0.59	0.58	0.58	0.57
Standard deviation	0.15	**0.16**	**0.16**	0.15	0.12	0.12	0.1	0.10	0.09	0.08

Table B.2: FRI scores obtained with *cat-ECM* with different values of β on the datasets.

B.1.3 Fuzzy Silhouette Index (FS)

	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Soybean	0.38	0.48	0.49	0.46	0.46	**0.5**	**0.5**	0.37	0.37	0.37
Zoo	0.42	0.55	0.57	0.55	0.49	0.52	0.52	**0.61**	0.57	0.5
Breast	0.11	0.19	0.20	0.15	0.15	0.15	**0.21**	0.12	0.15	0.15
Lung	0.11	0.11	0.09	0.08	NaN	0.05	0.03	0.12	0.11	**0.18**
Credits	0.27	0.25	0.33	0.33	0.31	0.33	0.33	**0.72**	0.33	**0.72**
Votes	0.54	0.57	0.58	0.58	0.59	0.6	0.6	0.6	0.61	**0.62**
Cars	0.12	0.12	0.14	0.14	0.15	0.15	0.15	**0.16**	**0.16**	0.15
Dermatology	0.15	0.12	0.15	0.19	0.07	0.21	**0.25**	0.09	NaN	0.06
Mushrooms	0.26	0.25	0.27	0.29	0.35	**0.36**	**0.36**	**0.36**	**0.36**	**0.36**
Mean	0.26	0.29	0.31	0.31	0.32	0.32	0.33	0.35	0.33	0.35
Standard deviation	0.16	0.19	0.19	0.19	0.19	0.19	0.19	**0.24**	0.19	0.23

Table B.3: FS scores obtained with *cat-ECM* with different values of β on the datasets.

B.1.4 Partition Coefficient (PC)

	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Soybean	**0.96**	0.92	0.86	0.77	0.67	0.58	0.52	0.46	0.42	0.40
Zoo	**0.89**	0.88	0.85	0.77	0.71	0.63	0.56	0.47	0.44	0.43
Breast	**0.74**	0.70	0.66	0.63	0.6	0.59	0.57	0.55	0.56	0.55
Lung	**0.72**	0.59	0.48	0.44	0.33	0.34	0.34	0.35	0.35	0.34
Credits	**0.96**	0.86	0.78	0.73	0.69	0.66	0.64	0.62	0.59	0.59
Votes	**0.98**	0.95	0.93	0.91	0.89	0.86	0.84	0.81	0.79	0.77
Cars	**0.78**	0.63	0.51	0.43	0.38	0.35	0.32	0.31	0.29	0.29
Dermatology	**0.71**	0.37	0.32	0.28	0.26	0.25	0.25	0.25	0.25	0.25
Mushrooms	**0.93**	0.78	0.73	0.69	0.66	0.63	0.61	0.59	0.58	0.57
Mean	0.85	0.74	0.68	0.63	0.58	0.54	0.52	0.49	0.47	0.47
Standard deviation	0.11	0.19	0.20	0.20	0.21	0.19	0.19	0.17	0.17	0.17

Table B.4: PC scores obtained with *cat-ECM* with different values of β on the datasets.

B.1.5 Partition Entropy (PE)

	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
Soybean	**0.07**	0.15	0.28	0.46	0.64	0.81	0.91	1.02	1.09	1.12
Zoo	**0.18**	0.21	0.3	0.50	0.64	0.8	0.96	1.18	1.25	1.27
Breast	**0.37**	0.44	0.5	0.54	0.58	0.59	0.61	0.63	0.62	0.63
Lung	0.48	0.69	0.87	0.93	**1.1**	**1.1**	1.09	1.08	1.08	1.09
Credits	**0.08**	0.24	0.35	0.42	0.47	0.51	0.54	0.56	0.59	0.59
Votes	**0.04**	0.08	0.12	0.15	0.19	0.23	0.27	0.3	0.34	0.37
Cars	**0.39**	0.69	0.92	1.05	1.15	1.21	1.26	1.28	1.31	1.32
Dermatology	**0.52**	1.16	1.26	1.32	1.37	1.38	1.38	1.38	1.38	1.38
Mushrooms	**0.12**	0.33	0.4	0.46	0.51	0.54	0.57	0.59	0.61	0.62
Mean	0.25	0.44	0.56	0.65	0.74	0.8	0.84	0.89	0.92	0.93
Standard deviation	0.19	0.35	0.38	0.37	0.38	0.37	0.37	0.38	0.38	0.38

Table B.5: PE scores obtained with *cat-ECM* with different values of β on the datasets.

B.2 Critical difference diagrams

B.2.1 Rand Index critical difference diagrams

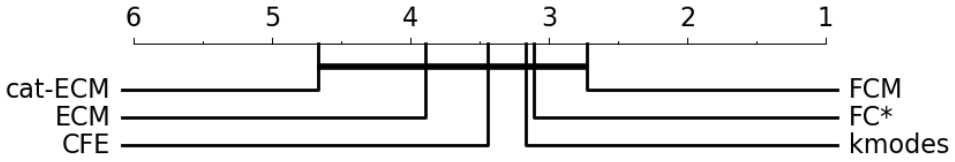

Figure B.1: Critical difference diagram obtained for RI scores with $\gamma = 0.1$ and $\beta = 2$.

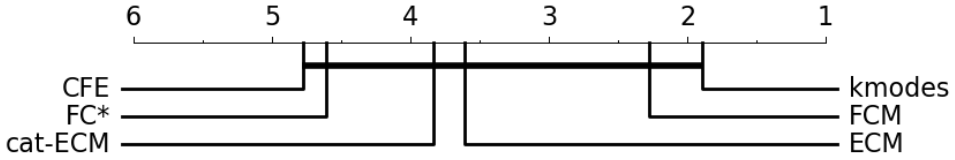

Figure B.2: Critical difference diagram obtained for RI scores with $\gamma = 0.1$ and $\beta = 1.1$.

B.2.2 Fuzzy Silhouette Index critical difference diagrams

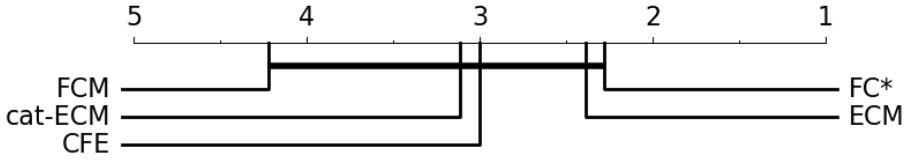

Figure B.3: Critical difference diagram obtained for FS scores with $\gamma = 0.1$ and $\beta = 1.1$.

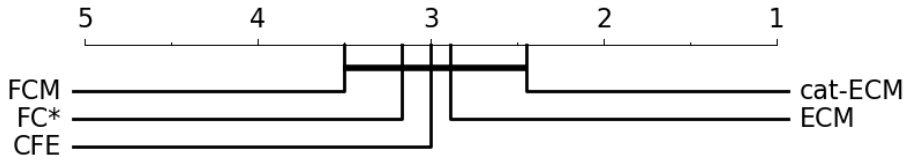

Figure B.4: Critical difference diagram obtained for FS scores with $\gamma = 0.1$ and $\beta = 2$.

B.3 Partitions comparisons

This section contains complement figures of the partition comparisons results in Chapter 4. The partitions are described in Table 4.1.

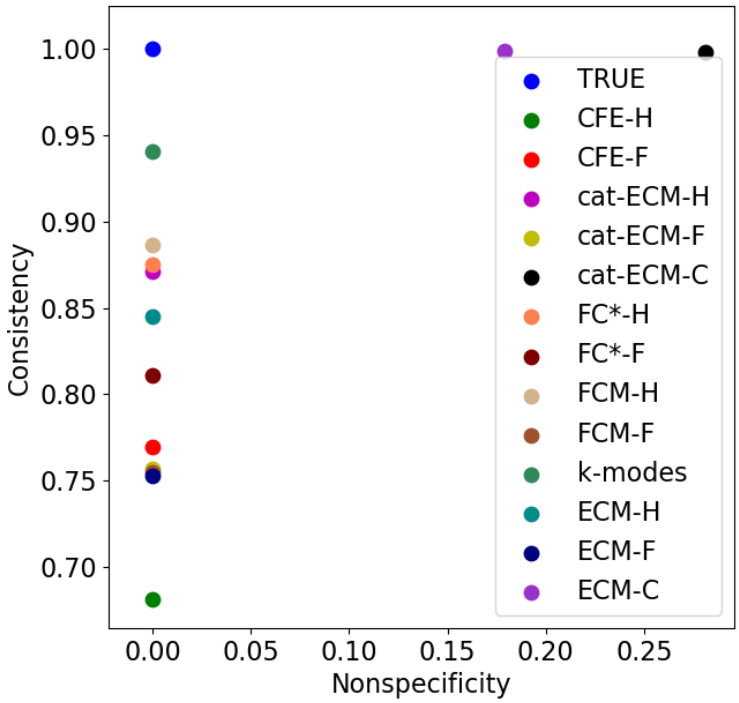

Figure B.5: Nonspecificity against Consistency obtained on the Zoo dataset with $\beta = 2$.

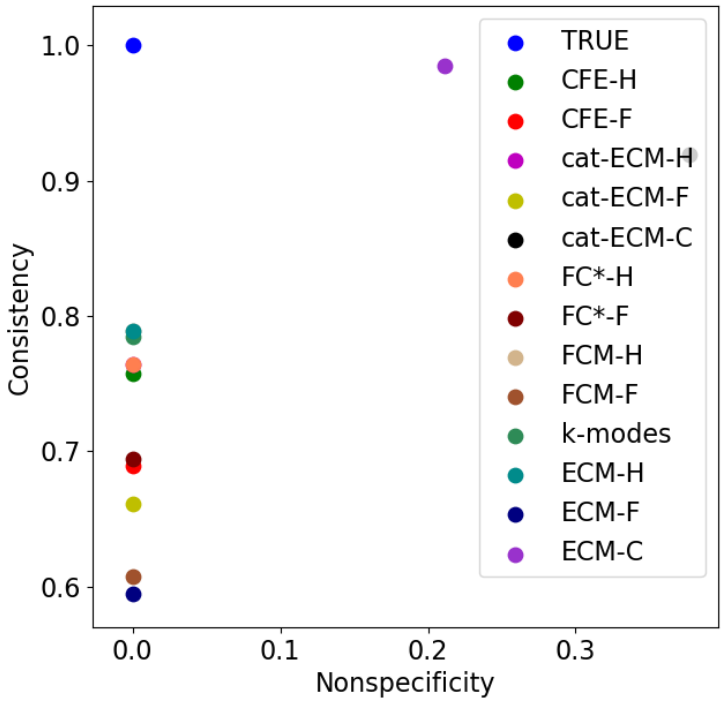

Figure B.6: Nonspecificity against consistency obtained on the Votes dataset with $\beta = 2$.

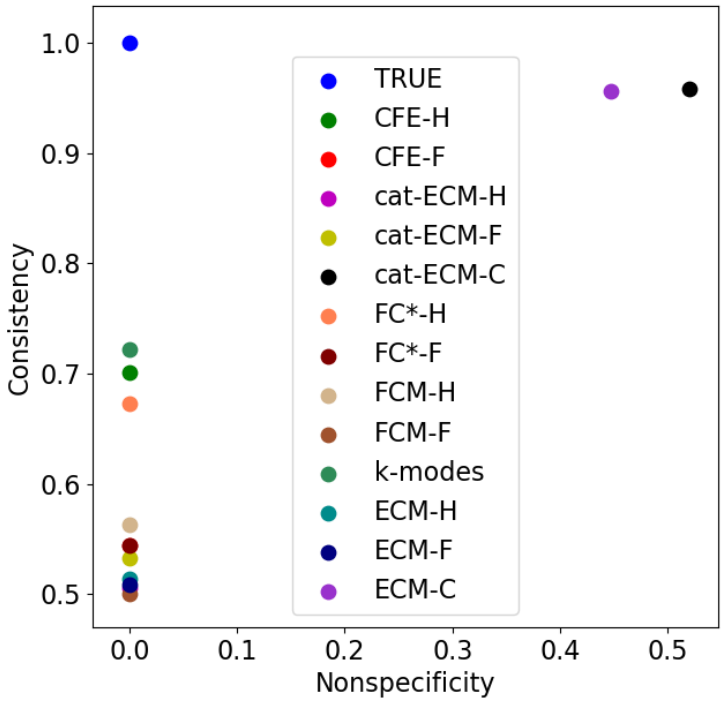

Figure B.7: Nonspecificity against consistency obtained on the Credits dataset with $\beta = 2$.

C — Complementary real-world applications results

In this appendix, supplement results of the experiments conducted in Chapter 5 are provided.

C.1 Descriptive statistics on the HRS dataset

The following table describes the most frequent attributes categories of the HRS dataset. Variables in this table are presented in Table 5.1.

	Cogn	1	2	3	4	5
Mild Activities	top	0	0	1	1	1
	freq	66	54	52	62	62
Read	top	1	1	1	1	1
	freq	71	80	80	86	88
Word Games	top	0	0	0	0	1
	freq	72	67	60	58	50
Use Computer	top	0	0	1	1	1
	freq	68	53	61	69	81
Sew/Knit	top	0	0	0	0	0
	freq	96	95	95	92	91
Walk 20min	top	0	1	1	1	1
	freq	53	50	51	54	56
Housing Problems	top	0	0	0	0	0
	freq	78	77	84	84	87
Drinks	top	0	1	1	1	1
	freq	52	53	64	65	72
Smokes	top	0	0	0	0	0
	freq	79	81	82	82	86

Table C.1: Frequencies in % of lifestyle factors categories.

C.2 Frequencies of cognitive categories in clusters

In this section, the relative and absolute frequencies of the five cognitive categories in each cluster obtained on the HRS dataset from *CFE* and *cat-ECM* are provided.

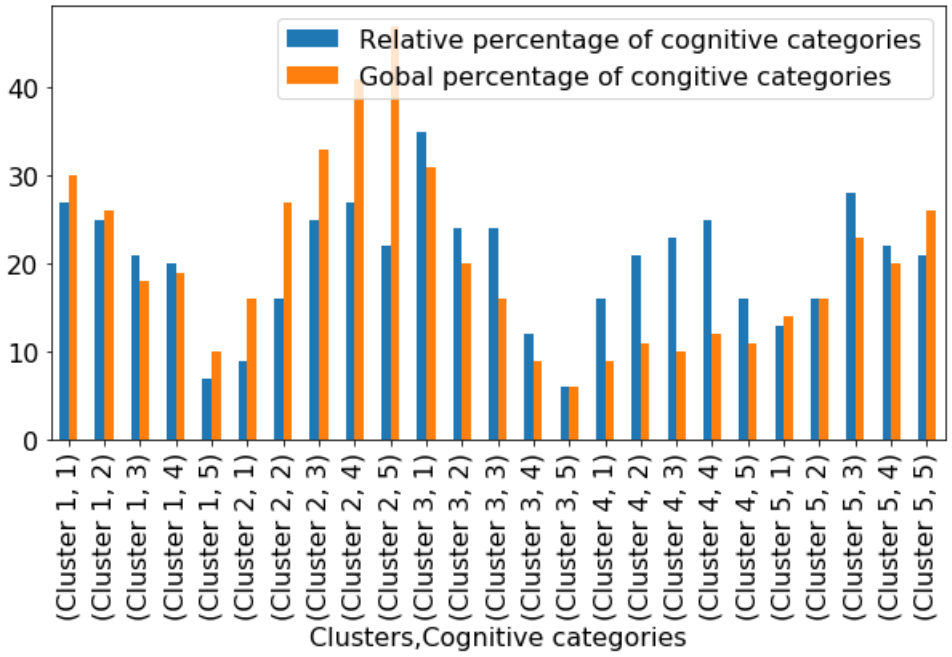

Figure C.1: Relative and absolute cognitive categories frequencies in clusters from *CFE*.

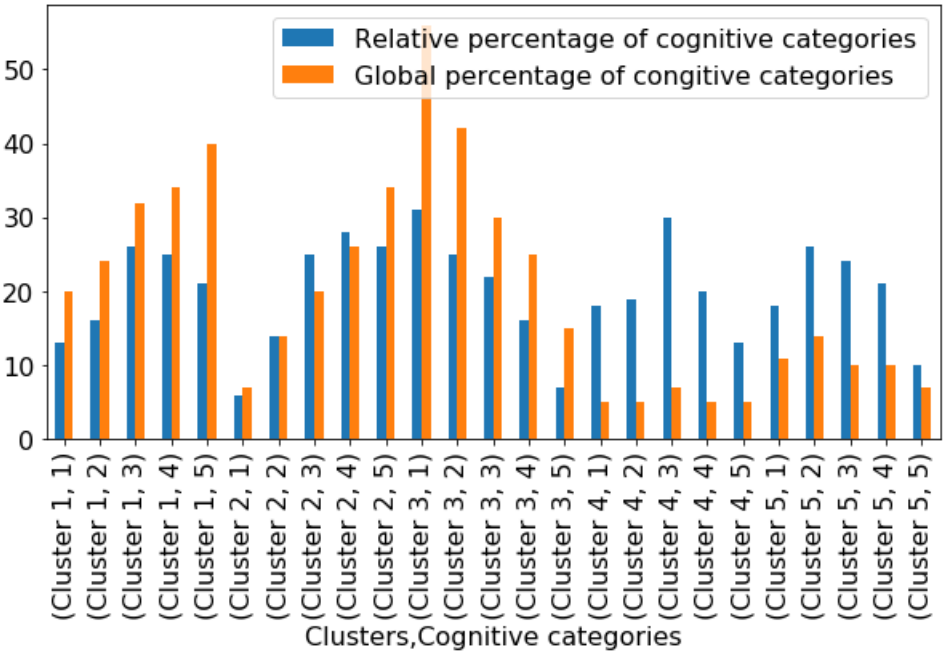

Figure C.2: Relative and absolute cognitive categories frequencies in clusters from *cat-ECM*.

C.3 Test values in each cognitive categories

Figure C.3 corresponds to the test values in each cognitive category. These values represent the expected test values if the is a 1-to-1 correspondence between the five clusters obtained from *CFE* and *cat-ECM* and the cognitive categories.

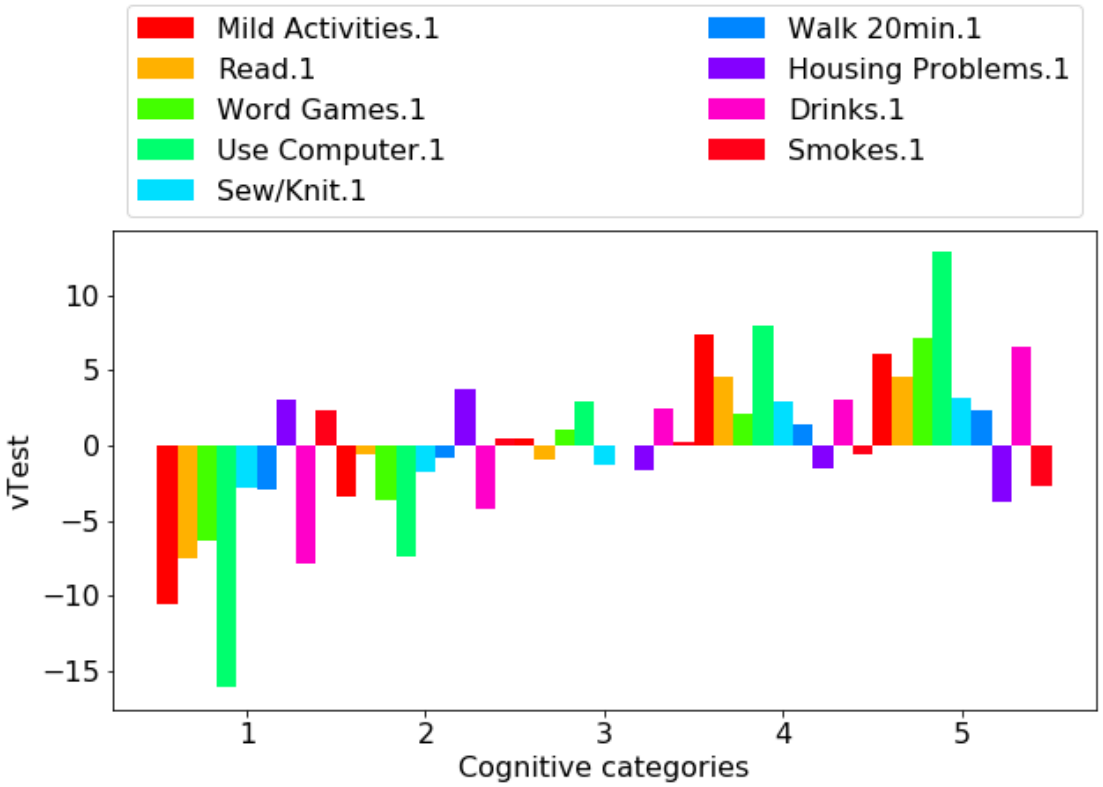

Figure C.3: vTests of lifestyle factors categories in each cognitive categories where the classes {12} and {45} correspond respectively the merged of cognitive categories 1 and 2 and 4 and 5. The class {3} corresponds to the cognitive category 3.

Bibliography

[1] E. A. Rodrigues, G. Christie, F. Farzan, and S. Moreno, "Does cognitive ageing follow an orchid and dandelion phenomenon?," *To be published*, 2019.

[2] H.-J. Zimmermann, *Fuzzy Set Theory — and Its Applications*, vol. 2001. Springer Netherlands, 01 2001.

[3] P. Bosc and H. Prade, "An introduction to the fuzzy set and possibility theory-based treatment of flexible queries and uncertain or imprecise databases," in *Uncertainty management in information systems*, pp. 285–324, Springer, 1997.

[4] M.-H. Masson, "Apports de la théorie des possibilités et des fonctions de croyance à l'analyse de données imprécises," 2005.

[5] D. Dubois and H. Prade, *Formal Representations of Uncertainty*, ch. 3, pp. 85–156. John Wiley Sons, Ltd, 2009.

[6] T. Denœux, D. Dubois, and H. Prade, "Representations of uncertainty in artificial intelligence: Probability and possibility," in *A Guided Tour of Artificial Intelligence Research*, pp. 69–117, Springer, 2020.

[7] L. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338 – 353, 1965.

[8] A. P. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *Ann. Math. Statist.*, vol. 38, pp. 325–339, 04 1967.

[9] G. Shafer, *A mathematical theory of evidence*, vol. 42. Princeton university press, 1976.

[10] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[11] L. A. Zadeh, "Fuzzy sets as a basis for a theory of possibility," *Fuzzy sets and systems*, vol. 1, no. 1, pp. 3–28, 1978.

[12] P. Walley and W. Peter, *Statistical Reasoning with Imprecise Probabilities*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis, 1991.

[13] C. Kahraman, B. Öztaysi, and S. C. Onar, "A comprehensive literature review of 50 years of fuzzy set theory," *International Journal of Computational Intelligence Systems*, vol. 9, pp. 3–24, 2016.

[14] A. Rényi *et al.*, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, The Regents of the University of California, 1961.

[15] C. Beck, "Generalised information and entropy measures in physics," *Contemporary Physics*, vol. 50, no. 4, pp. 495–510, 2009.

[16] A. B. Templeman, "Entropy and civil engineering optimization," in *Optimization and Artificial Intelligence in Civil and Structural Engineering*, pp. 87–105, Springer, 1992.

[17] R. Zhou, R. Cai, and G. Tong, "Applications of entropy in finance: A review," *Entropy*, vol. 15, no. 11, pp. 4909–4931, 2013.

[18] C. Adami, "Information theory in molecular biology," *Physics of Life Reviews*, vol. 1, no. 1, pp. 3–22, 2004.

[19] K. D. Bailey, *Social entropy theory*. SUNY Press, 1990.

[20] V. P. Singh, *Entropy theory and its application in environmental and water engineering*. John Wiley & Sons, 2013.

[21] F. E. Ruiz, P. S. Pérez, and B. I. Bonev, *Information theory in computer vision and pattern recognition*. Springer Science & Business Media, 2009.

[22] D. Klaua, "Ein ansatz zur mehrwertigen mengenlehre," *Mathematische Nachrichten*, vol. 33, no. 5-6, pp. 273–296, 1967.

[23] E. P. Klement, R. Mesiar, and E. Pap, *Triangular norms*, vol. 8. Springer Science & Business Media, 2013.

[24] H.-J. Zimmermann, "Fuzzy set theory," *WIREs Computational Statistics*, vol. 2, no. 3, pp. 317–332, 2010.

[25] G. Bojadziev and M. Bojadziev, *Fuzzy Sets, Fuzzy Logic, Applications*. WORLD SCIENTIFIC, 1996.

[26] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer US, 1981.

[27] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[28] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[29] J. Macqueen, "Some methods for classification and analysis of multivariate observations," in *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

[30] D. Kim, K. Lee, and D. Lee, "Fuzzy clustering of categorical data using fuzzy centroids," *Pattern recognition letters*, vol. 25, no. 11, pp. 1263–1271, 2004.

[31] P. Smets and R. Kennes, "The transferable belief model," *Artificial Intelligence*, vol. 66, no. 2, pp. 191 – 234, 1994.

[32] G. Shafer, "Perspectives on the theory and practice of belief functions," *International Journal of Approximate Reasoning*, vol. 4, no. 5, pp. 323–362, 1990.

[33] P. Smets, "Constructing the pignistic probability function in a context of uncertainty.," in *UAI*, vol. 89, pp. 29–40, 1989.

[34] B. R. Cobb and P. P. Shenoy, "On the plausibility transformation method for translating belief function models to probability models," *International journal of approximate reasoning*, vol. 41, no. 3, pp. 314–330, 2006.

[35] P. Smets, "Decision making in the tbm: the necessity of the pignistic transformation," *International journal of approximate reasoning*, vol. 38, no. 2, pp. 133–147, 2005.

[36] A.-L. Jousselme, C. Liu, D. Grenier, and É. Bossé, "Measuring ambiguity in the evidence theory," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 36, no. 5, pp. 890–903, 2006.

[37] D. Harmanec and G. J. Klir, "Measuring total uncertainty in dempster-shafer theory: A novel approach," *International journal of general system*, vol. 22, no. 4, pp. 405–419, 1994.

[38] G. J. Klir, M. J. Wierman, and J. Kacprzyk, *Uncertainty-Based Information: Elements of Generalized Information Theory*. Physica-Verlag, 2nd ed., 1999.

[39] R. V. Hartley, "Transmission of information 1," *Bell System technical journal*, vol. 7, no. 3, pp. 535–563, 1928.

[40] R. Körner and W. Näther, "On the specificity of evidences," *Fuzzy sets and systems*, vol. 71, no. 2, pp. 183–196, 1995.

[41] Y. Yang, D. Han, and J. Dezert, "A new non-specificity measure in evidence theory based on belief intervals," *Chinese Journal of Aeronautics*, vol. 29, no. 3, pp. 704–713, 2016.

[42] Y. Deng, "Uncertainty measure in evidence theory," *Science China Information Sciences*, vol. 63, no. 11, pp. 1–19, 2020.

[43] T. Denoeux, "A k-nearest neighbor classification rule based on dempster-shafer theory," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 25, no. 5, pp. 804–813, 1995.

[44] T. Denœux, "Logistic regression, neural networks and dempster–shafer theory: A new perspective," *Knowledge-Based Systems*, vol. 176, pp. 54–67, 2019.

[45] Y. Zhang, Q. Zeng, Y. Liu, and B. Shen, "Integrated data fusion using dempster-shafer theory," in *2015 First International Conference on Computational Intelligence Theory, Systems and Applications (CCITSA)*, pp. 98–103, 2015.

[46] M. Rombaut and V. Cherfaoui, "Decision making in data fusion using dempster-shafer's theory," *IFAC Proceedings Volumes*, vol. 30, no. 7, pp. 339–343, 1997. 3rd IFAC Symposium on Intelligent Components and Instruments For Control Applications 1997 (SICICA '97), Annecy, France, 9-11 June.

[47] P. Shenoy, *Using Dempster-Shafer's Belief-Function Theory in Expert Systems*, pp. 395–. 01 1994.

[48] J. Guan, D. A. Bell, J. Pavlin, and V. R. Lesser, "Dempster-shafer theory and rule strengths in expert systems," in *IEE Colloquium on Reasoning Under Uncertainty*, pp. 6/1–6/3, 1990.

[49] U. RAKOWSKY, "Fundamentals of the dempster-shafer theory and its applications to reliability modeling," *International Journal of Reliability, Quality and Safety Engineering - IJRQSE*, vol. 14, 12 2007.

[50] J. A. Malpica, M. C. Alonso, and M. A. Sanz, "Dempster–shafer theory in geographic information systems: A survey," *Expert Systems with Applications*, vol. 32, no. 1, pp. 47–55, 2007.

[51] M.-H. Masson and T. Denœux, "Ecm: An evidential version of the fuzzy c-means algorithm," *Pattern Recognition*, vol. 41, pp. 1384–1397, 04 2008.

[52] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[53] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017.

[54] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE transactions on fuzzy systems*, vol. 1, no. 2, pp. 98–110, 1993.

[55] P. Maji and S. K. Pal, "Rough set based generalized fuzzy c-means algorithm and quantitative indices," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 6, pp. 1529–1540, 2007.

[56] T. Denoeux and O. Kanjanatarakul, "Evidential clustering: a review," in *International symposium on integrated uncertainty in knowledge modelling and decision making*, pp. 24–35, Springer, 2016.

[57] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.

[58] J. F. Kolen and T. Hutcheson, "Reducing the time complexity of the fuzzy c-means algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 10, no. 2, pp. 263–267, 2002.

[59] M. Masson and T. Denœux, "ECM: An evidential version of the fuzzy c-means algorithm," *Pattern Recognition*, vol. 41, no. 4, pp. 1384–1397, 2008.

[60] Sumit Sen and R. N. Dave, "Clustering of relational data containing noise and outliers," in *1998 IEEE International Conference on Fuzzy Systems Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36228)*, vol. 2, pp. 1411–1416 vol.2, 1998.

[61] D. Tran and M. Wagner, "Fuzzy entropy clustering," in *Ninth IEEE International Conference on Fuzzy Systems. FUZZ- IEEE 2000*, vol. 1, pp. 152–157 vol.1, May 2000.

[62] M. Zarinbal, M. Fazel Zarandi, and I. Turksen, "Relative entropy fuzzy c-means clustering," *Information Sciences*, vol. 260, pp. 74–97, 2014.

[63] A. Lorette, X. Descombes, and J. Zerubia, "Fully unsupervised fuzzy clustering with entropy criterion," in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol. 3, pp. 986–989 vol.3, Sep. 2000.

[64] J. Yao, M. Dash, S. Tan, and H. Liu, "Entropy-based fuzzy clustering and fuzzy modeling," *Fuzzy Sets and Systems*, vol. 113, no. 3, pp. 381 – 388, 2000.

[65] N. B. Karayiannis, "Meca: maximum entropy clustering algorithm," in *Proceedings of 1994 IEEE 3rd International Fuzzy Systems Conference*, pp. 630–635 vol.1, June 1994.

[66] J. C. Xavier, A. M. P. Canuto, N. D. Almeida, and L. M. G. Gonçalves, "A comparative analysis of dissimilarity measures for clustering categorical data," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2013.

[67] R. W. Hamming, "Error detecting and error correcting codes," *The Bell System Technical Journal*, vol. 29, no. 2, pp. 147–160, 1950.

[68] Z. Huang and M. Ng, "A fuzzy k-modes algorithm for clustering categorical data," *IEEE Transactions on Fuzzy Systems*, vol. 7, no. 4, pp. 446–452, 1999.

[69] W. Wang and Y. Zhang, "On fuzzy cluster validity indices," *Fuzzy Sets and Systems*, vol. 158, no. 19, pp. 2095–2117, 2007. Theme: Data Analysis.

[70] J. Bezdek, "Cluster validity with fuzzy sets," *Journal of Cybernetics*, vol. 3, 07 1973.

[71] J. C. Bezdek, "Numerical taxonomy with fuzzy sets," *Journal of mathematical biology*, vol. 1, no. 1, pp. 57–71, 1974.

[72] R. Campello and E. Hruschka, "A fuzzy extension of the silhouette width criterion for cluster analysis," *Fuzzy Sets and Systems*, vol. 157, no. 21, pp. 2858–2875, 2006.

[73] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[74] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[75] R. Campello, "A fuzzy extension of the rand index and other related indexes for clustering and classification assessment," *Pattern Recognition Letters*, vol. 28, no. 7, pp. 833–841, 2007.

[76] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," *Expert Systems with Applications*, vol. 40, no. 1, pp. 200–210, 2013.

[77] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 01 2006.

[78] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937.

[79] M. Friedman, "A Comparison of Alternative Tests of Significance for the Problem of m Rankings," *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86 – 92, 1940.

[80] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in statistics*, pp. 196–202, Springer, 1992.

[81] N. J. de Vos, "kmodes categorical clustering library." https://github.com/nicodv/kmodes, 2015–2021.

[82] J. Warner, J. Sexauer, scikit fuzzy, twmeggs, alexsavio, A. Unnikrishnan, G. Castelão, F. A. Pontes, T. Uelwer, pd2f, laurazh, F. Batista, alexbuy, W. V. den Broeck, W. Song, T. G. Badger, R. A. M. Pérez, J. F. Power, H. Mishra, G. O. Trullols, A. Hörteborn, and 99991, "Jdwarner/scikit-fuzzy: Scikit-fuzzy version 0.4.2," Nov. 2019.

[83] M. B. Ferraro, P. Giordani, and A. Serafini, "fclust: An R Package for Fuzzy Clustering," *The R Journal*, vol. 11, no. 1, pp. 198–210, 2019.

[84] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data Mining and Knowledge Discovery*, vol. 33, no. 4, pp. 917–963, 2019.

[85] K. Zhou, C. Fu, and S. Yang, "Fuzziness parameter selection in fuzzy c-means: the perspective of cluster validation," *Science China Information Sciences*, vol. 57, no. 11, pp. 1–8, 2014.

[86] K.-L. Wu, "Analysis of parameter selections for fuzzy c-means," *Pattern Recognition*, vol. 45, pp. 407–415, 01 2012.

[87] N. Pal and J. Bezdek, "On cluster validity for the fuzzy c-means model," *IEEE Trans. Fuzzy Syst.*, vol. 3, pp. 370–379, 1995.

[88] A. Lorette, X. Descombes, and J. Zerubia, "Fully unsupervised fuzzy clustering with entropy criterion," in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol. 3, pp. 986–989, IEEE, 2000.

[89] T. Denœux, S. Li, and S. Sriboonchitta, "Evaluating and comparing soft partitions: An approach based on dempster–shafer theory," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 3, pp. 1231–1244, 2018.

[90] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE transactions on fuzzy systems*, vol. 13, no. 4, pp. 517–530, 2005.

[91] A.-L. Jousselme, D. Grenier, and Bossé, "A new distance between two bodies of evidence," *Information Fusion*, vol. 2, pp. 91–101, 06 2001.

[92] T. Denœux, "Inner and outer approximation of belief structures using a hierarchical clustering approach," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 9, pp. 437–460, 08 2001.

[93] R. N. Dave, "Characterization and detection of noise in clustering," *Pattern Recognition Letters*, vol. 12, no. 11, pp. 657–664, 1991.

[94] Alzheimer's Disease International, Andres Wimo Gemma-Claire Ali, Maëlenn Guerchet, Martin Prince, Matthew Prina, Yu-Tzu Wu, "World alzheimer report 2015, the global impact of dementia: An analysis of prevalence, incidence, cost and trends," 2015.

[95] U. Nations, "World population ageing 2019: highlights (st/esa/ser. a/430)," 2019.

[96] W. H. Organization *et al.*, "Risk reduction of cognitive decline and dementia: Who guidelines," 2019.

[97] M. Baumgart, H. M. Snyder, M. C. Carrillo, S. Fazio, H. Kim, and H. Johns, "Summary of the evidence on modifiable risk factors for cognitive decline and dementia: A population-based perspective," *Alzheimer's Dementia*, vol. 11, no. 6, pp. 718–726, 2015.

[98] G. J. Christie, T. Hamilton, B. D. Manor, N. A. Farb, F. Farzan, A. Sixsmith, J.-J. Temprado, and S. Moreno, "Do lifestyle activities protect against cognitive decline in aging? a review," *Frontiers in aging neuroscience*, vol. 9, p. 381, 2017.

[99] S. A. Small, "Age-Related Memory Decline: Current Concepts and Future Directions," *Archives of Neurology*, vol. 58, pp. 360–364, 03 2001.

[100] H. Slater and J. Young, "A review of brief cognitive assessment tests," *Reviews in Clinical Gerontology*, vol. 23, no. 2, p. 164–176, 2013.

[101] R. S. Wilson, E. Segawa, P. A. Boyle, and D. A. Bennett, "Influence of late-life cognitive activity on cognitive health," *Neurology*, vol. 78, no. 15, pp. 1123–1129, 2012.

[102] D. A. Bennett, J. A. Schneider, A. S. Buchman, C. M. de Leon, J. L. Bienias, and R. S. Wilson, "The rush memory and aging project: study design and baseline characteristics of the study cohort," *Neuroepidemiology*, vol. 25, no. 4, pp. 163–175, 2005.

[103] M. F. Folstein, S. E. Folstein, and P. R. McHugh, ""mini-mental state": a practical method for grading the cognitive state of patients for the clinician," *Journal of psychiatric research*, vol. 12, no. 3, pp. 189–198, 1975.

[104] K. A. Bollen and P. J. Curran, *Latent curve models: A structural equation perspective*, vol. 467. John Wiley & Sons, 2006.

[105] L. M. Collins and A. G. Sayer, *New methods for the analysis of change.* American Psychological Association, 2001.

[106] T. Fritsch, M. J. McClendon, K. A. Smyth, A. J. Lerner, R. P. Friedland, and J. D. Larsen, "Cognitive functioning in healthy aging: the role of reserve and lifestyle factors early in life," *The Gerontologist*, vol. 47, no. 3, pp. 307–322, 2007.

[107] T. Fritsch, K. A. Smyth, M. J. McClendon, P. K. Ogrocki, C. Santillan, J. D. Larsen, and M. E. Strauss, "Associations between dementia/mild cognitive impairment and cognitive performance and activity levels in youth," *Journal of the American Geriatrics Society*, vol. 53, no. 7, pp. 1191–1196, 2005.

[108] K. A. Welsh, J. C. Breitner, and K. M. Magruder-Habib, "Detection of dementia in the elderly using telephone screening of cognitive status.," *Neuropsychiatry, Neuropsychology, & Behavioral Neurology*, 1993.

[109] M. McClendon, *Multiple Regression and Causal Analysis.* Waveland Press, 2002.

[110] E. Pedhazur, *Multiple Regression in Behavioral Research: Explanation and Prediction.* Harcourt Brace College Publishers, 1997.

[111] S. H. Lee and Y. B. Kim, "Which type of social activities may reduce cognitive decline in the elderly?: a longitudinal population-based study," *BMC geriatrics*, vol. 16, no. 1, pp. 1–9, 2016.

[112] Y. Kang, D. L. Na, and S. Hahn, "A validity study on the korean mini-mental state examination (k-mmse) in dementia patients," *Journal of the Korean neurological association*, vol. 15, no. 2, pp. 300–308, 1997.

[113] R. Brant, "Assessing proportionality in the proportional odds model for ordinal logistic regression," *Biometrics*, pp. 1171–1178, 1990.

[114] W. T. Boyce, M. Chesney, A. Alkon, J. M. Tschann, S. Adams, B. Chesterman, F. Cohen, P. Kaiser, S. Folkman, and D. Wara, "Psychobiologic reactivity to stress and childhood respiratory illnesses: Results of two prospective studies," *Psychosomatic medicine*, vol. 57, no. 5, pp. 411–422, 1995.

[115] J. Belsky, "Variation in susceptibility to environmental influence: An evolutionary argument," *Psychological inquiry*, vol. 8, no. 3, pp. 182–186, 1997.

[116] C. S. Barr, T. K. Newman, C. Shannon, C. Parker, R. L. Dvoskin, M. L. Becker, M. Schwandt, M. Champoux, K. P. Lesch, D. Goldman, *et al.*, "Rearing condition and rh5-httlpr interact to influence limbic-hypothalamic-pituitary-adrenal axis response to stress in infant macaques," *Biological psychiatry*, vol. 55, no. 7, pp. 733–738, 2004.

[117] W. T. Boyce, "Differential susceptibility of the developing brain to contextual adversity and stress," *Neuropsychopharmacology*, vol. 41, no. 1, pp. 142–162, 2016.

[118] W. T. Boyce, P. Levitt, F. D. Martinez, B. S. McEwen, and J. P. Shonkoff, "Genes, environments, and time: the biology of adversity and resilience," *Pediatrics*, vol. 147, no. 2, 2021.

[119] J. Belsky and M. Pluess, "Genetic moderation of early child-care effects on social functioning across childhood: A developmental analysis," *Child Development*, vol. 84, no. 4, pp. 1209–1225, 2013.

[120] D. Berry, K. Deater-Deckard, K. McCartney, Z. Wang, and S. A. Petrill, "Gene–environment interaction between dopamine receptor d4 7-repeat polymorphism and early maternal sensitivity predicts inattention trajectories across middle childhood," *Development and psychopathology*, vol. 25, no. 2, pp. 291–306, 2013.

[121] D. A. Windhorst, V. R. Mileva-Seitz, M. Linting, A. Hofman, V. W. Jaddoe, F. C. Verhulst, H. Tiemeier, M. H. van IJzendoorn, and M. J. Bakermans-Kranenburg, "Differential susceptibility in a developmental perspective: Drd4 and maternal sensitivity predicting externalizing behavior," *Developmental psychobiology*, vol. 57, no. 1, pp. 35–49, 2015.

[122] A. Yardimci, "Soft computing in medicine," *Applied Soft Computing*, vol. 9, no. 3, pp. 1029–1043, 2009.

[123] R. Alizadehsani, M. Roshanzamir, S. Hussain, A. Khosravi, A. Koohestani, M. H. Zangooei, M. Abdar, A. Beykikhoshk, A. Shoeibi, A. Zare, M. Panahiazar, S. Nahavandi, D. Srinivasan, A. F.

Atiya, and U. R. Acharya, "Handling of uncertainty in medical data using machine learning and probability theory techniques: A review of 30 years (1991-2020)," *CoRR*, vol. abs/2008.10114, 2020.

[124] M. F. Abbod, D. G. von Keyserlingk, D. A. Linkens, and M. Mahfouf, "Survey of utilisation of fuzzy technology in medicine and healthcare," *Fuzzy Sets and Systems*, vol. 120, no. 2, pp. 331–349, 2001.

[125] Y. Sekita and Y. Tabata, "A health status index model using a fuzzy approach," *European Journal of Operational Research*, vol. 3, no. 1, pp. 40–49, 1979.

[126] T. Sedbrook, H. Wright, and R. Wright, "A visual fuzzy cluster system for patient analysis," *Medical Informatics*, vol. 18, no. 4, pp. 321–329, 1993.

[127] P. P. Shenoy, "Using Dempster-Shafer's belief-function theory in expert systems," in *Applications of Artificial Intelligence X: Knowledge-Based Systems* (G. Biswas, ed.), vol. 1707, pp. 2 – 14, International Society for Optics and Photonics, SPIE, 1992.

[128] S. Peñafiel, N. Baloian, J. A. Pino, J. Quinteros, Riquelme, H. Sanson, and D. Teoh, "Associating risks of getting strokes with data from health checkup records using dempster-shafer theory," in *2018 20th International Conference on Advanced Communication Technology (ICACT)*, pp. 1–1, 2018.

[129] A. Sagdoldanova, L. Atymtayeva, and Z. Yespolayeva, "Medicine recommendation technique by using dempster-shafer theory," *Adv. Eng. Tec. Appl*, vol. 6, no. 3, pp. 27–32, 2017.

[130] M. Raza, I. Gondal, D. Green, and R. L. Coppel, "Classifier fusion using dempster-shafer theory of evidence to predict breast cancer tumors," in *TENCON 2006 - 2006 IEEE Region 10 Conference*, pp. 1–4, 2006.

[131] H. Huang, Z. Yan, Y. Chen, and F. Liu, "A social contagious model of the obesity epidemic," *Scientific reports*, vol. 6, no. 1, pp. 1–9, 2016.

[132] J. K. Kim, M. J. Choi, J. S. Lee, J. H. Hong, C.-S. Kim, S. I. Seo, C. W. Jeong, S.-S. Byun, K. C. Koo, B. H. Chung, *et al.*, "A deep belief network and dempster-shafer-based multiclassifier for the pathology stage of prostate cancer," *Journal of healthcare engineering*, vol. 2018, 2018.

[133] Y. He, M. Y. Hussaini, Y. U. Gong, and Y. Xiao, "Optimal unified combination rule in application of dempster-shafer theory to lung can-

cer radiotherapy dose response outcome analysis," *Journal of applied clinical medical physics*, vol. 17, no. 1, pp. 4–11, 2016.

[134] W. Chen, Y. Cui, Y. He, Y. Yu, J. Galvin, Y. M. Hussaini, and Y. Xiao, "Application of dempster–shafer theory in dose response outcome analysis," *Physics in Medicine & Biology*, vol. 57, no. 17, p. 5575, 2012.

[135] Z. Li, G. Wen, and N. Xie, "An approach to fuzzy soft sets in decision making based on grey relational analysis and dempster–shafer theory of evidence: An application in medical diagnosis," *Artificial intelligence in medicine*, vol. 64, no. 3, pp. 161–171, 2015.

[136] J.-Y. Shi, X.-Q. Shang, K. Gao, S.-W. Zhang, and S.-M. Yiu, "An integrated local classification model of predicting drug-drug interactions via dempster-shafer theory of evidence," *Scientific reports*, vol. 8, no. 1, pp. 1–11, 2018.

[137] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 42, no. 2, pp. 109–127, 1980.

[138] A. Morineau, "Note sur la caractérisation statistique d'une classe et les valeurs-tests," *Bulletin Technique Centre Statistique Informatique Appliquées*, vol. 2, no. 1-2, pp. 20–27, 1984.

[139] F. Lionetti, A. Aron, E. N. Aron, G. L. Burns, J. Jagiellowicz, and M. Pluess, "Dandelions, tulips and orchids: Evidence for the existence of low-sensitive, medium-sensitive and high-sensitive individuals," *Translational psychiatry*, vol. 8, no. 1, pp. 1–11, 2018.

[140] O. Krakovska, G. Christie, A. Sixsmith, M. Ester, and S. Moreno, "Performance comparison of linear and non-linear feature selection methods for the analysis of large survey datasets," *Plos one*, vol. 14, no. 3, p. e0213584, 2019.

[141] J. Cohen, *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[142] M. B. Ferraro and P. Giordani, "A review and proposal of (fuzzy) clustering for nonlinearly separable data," *International Journal of Approximate Reasoning*, vol. 115, pp. 13–31, 2019.

[143] M. García-Magariños and J. A. Vilar, "A framework for dissimilarity-based partitioning clustering of categorical time series," *Data mining and knowledge discovery*, vol. 29, no. 2, pp. 466–502, 2015.